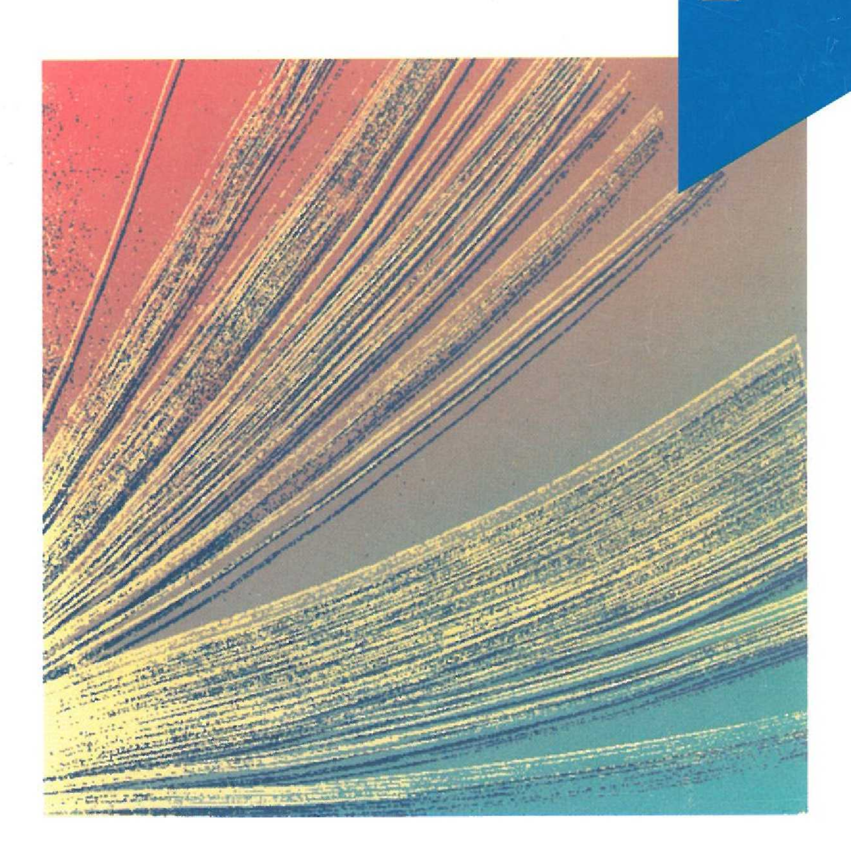
INSEE MÉTHODES
N° 84-85-86

INSEE MÉTHODES

ACTES DES JOURNÉES DE MÉTHODOLOGIE STATISTIQUE

17 et 18 mars 1998





ACTES DES JOURNÉES DE MÉTHODOLOGIE STATISTIQUE

17 et 18 mars 1998

Les résultats contenus dans ce livre sont
le fruit d'un travail de recherche.
Il n'engagent que leurs auteurs.

RÉPUBLIQUE FRANÇAISE
INSTITUT NATIONAL
DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Direction Générale
18, boulevard Adolphe-Pinard - 75675 Paris cedex 14

Directeur de la publication : Paul Champsaur
Coordinateur de la rédaction : Stéphane Tagnani
Maquettistes : Patricia Landais et Thérèse Pécheux

SOMMAIRE

SESSION 3 : LE LOGICIEL DE CALCUL DE PRÉCISION POULPE

SESSION 4 : LES INDICES

SESSION 5 : FAIRE POUR APPRENDRE : TROIS EXPÉRIENCES DE FORMATION ACTIVE AUX ENQUÊTES

PRESENTATION

Ketty Attal-Toubert
Unité méthodes statistiques - Insee

Organisées par l'unité Méthodes statistiques de l'Insee, les sixièmes Journées de méthodologie statistique[1] *se sont tenues les 17 et 18 mars 1998 à Paris, au Centre de conférence Pierre Mendès France du ministère de l'Economie, des Finances et de l'Industrie. Le succès de ce rendez-vous annuel ne se dément pas : au total, environ 400 personnes ont assisté aux différentes conférences et sessions thématiques.*

Les questions relatives à la précision des estimations ont occupé une place importante : présentation du logiciel POULPE de calcul de précision des enquêtes, exposé comparatif, par un confrère de Statistique Canada, de différentes méthodes d'estimation de la variance en présence d'imputation de la non-réponse. Les nouvelles orientations relatives aux recensements de population ont été débattues, et un point de vue américain a été donné par un collègue du Bureau of the Census des Etats-Unis. Il a également été question de la méthodologie des enquêtes par panel, avec en exemple le panel européen de ménages, de la collecte par enquêteurs et des indices de prix. Enfin, des enseignants, parmi lesquels un professeur ivoirien, ont rendu compte d'expériences de formation active aux enquêtes.

La conférence inaugurale de ces sixièmes Journées a porté sur les préoccupations européennes en matière d'harmonisation et de comparabilité.

Tant ses producteurs que ses utilisateurs sont de plus en plus attentifs à la qualité de l'information statistique. Dans la conférence inaugurale, Daniel Defays (Eurostat) a expliqué que dans le contexte actuel de globalisation économique et d'intégration européenne, la comparabilité des données, au-delà de leur précision intrinsèque, était une composante essentielle de la qualité de l'information statistique. Une harmonisation complète des méthodes de mesure ne paraît toutefois pas envisageable, les spécificités des pays exigeant souvent des méthodologies adaptées. En matière de comparabilité, les progrès doivent donc résider dans une meilleure compréhension des différentes sources, dans la construction d'une théorie de l'erreur qui permette d'identifier les écarts liés aux méthodes, et dans l'utilisation de modèles d'analyse des résultats permettant de « purifier » les mesures.

---

1. Les actes des cinq premières Journées sont parus dans la collection *Insee Méthodes*.

Recensements de la population

Lors de la première conférence spéciale, Yves Thibaudeau (Bureau of the Census) a exposé la stratégie américaine, à savoir économie et efficacité, pour le recensement des Etats-Unis de l'an 2000 : développement de partenariats avec des institutions locales et des services de marketing, simplification des opérations et des questionnaires, adoption d'une technologie intelligente (lecture optique par exemple) et d'une méthodologie statistique permettant de corriger les biais liés au sous-dénombrement et à la non-réponse. Le biais de sous-dénombrement devrait être estimé à l'aide d'une technique de « capture-recapture », la correction du biais de non-réponse s'appuiera sur une enquête auprès d'un échantillon de non-répondants.

Dans un exposé introductif au débat sur l'avenir des recensements, Michel Isnard (Insee-DG, département de la démographie) a présenté un projet de recensement en continu, préparé par une mission d'étude sur l'avenir des statistiques de population. Les difficultés que pose un recensement général (coût, adéquation du personnel de l'Insee, acceptation par la population) ont en effet incité l'Institut à réfléchir à une nouvelle forme de recensement tournant. Différents thèmes ont été abordés, comme la définition de la population légale, l'organisation de la collecte, les liens avec le recensement de 1999, la méthodologie d'échantillonnage et la diffusion.

Le débat a donné lieu à des échanges sur les grandes transformations qu'apporterait le passage à un recensement en continu concernant la nature des statistiques produites, leur fréquence, leur précision, les modalités de diffusion...

Estimation de variance en présence d'imputation

Une deuxième conférence spéciale a été consacrée à l'estimation de variance en présence d'imputation de la non-réponse. En effet, l'imputation des non-réponses peut non seulement introduire un biais dans les estimations, mais aussi fausser le calcul de la variance. Eric Rancourt (Statistique Canada) a présenté puis comparé différentes méthodes : imputation multiple, approche assistée d'un modèle, approche en deux phases, technique du Jackknife, Bootstrap, méthode pour hot-deck, méthode d'imputation de tous les cas, et méthode des échantillons balancés répétés. La méthode assistée d'un modèle est incorporée à un logiciel informatique en cours de développement par Statistique Canada.

Le panel européen de ménages

Jean-Claude Deville (Insee-DG, unité Méthodes statistiques) a ouvert cette session par un exposé des contraintes logiques relatives à la construction des données de

panel, contraintes liées à la nature des unités, à leur identification, à la façon de les échantillonner, aux techniques de pondération, à la correction de la non-réponse. Il a introduit un formalisme unificateur des différentes formes de collecte dans le but de rendre accessibles les analyses habituelles, notamment le calcul de variance.

Le panel européen de ménages[2] est une enquête réalisée depuis 1994 dans l'ensemble des pays de l'Union européenne, à objet d'observer l'évolution de la situation d'activité et de revenus des individus des ménages. Dominique Ansieau (Insee, direction régionale de Lorraine) s'est appuyé sur l'exemple de cette enquête pour évoquer le problème général de l'attrition des panels, et détailler un certain nombre de dispositions pratiques permettant d'atténuer le phénomène : amélioration de l'identification et du suivi (cas de déménagements) des ménages et des personnes, remise de cadeaux aux ménages répondants, réalisation des enquêtes successives auprès d'un même ménage par le même enquêteur.

Christine Chambaz et Nadine Legendre (Insee-DG, division Revenus et patrimoines des ménages) se sont penchées sur l'étude et la correction de la non-réponse dans le panel européen de ménages : analyse des caractéristiques des non-répondants, et correction de la non-réponse par post-stratification. Pour le calcul des pondérations, ont été éprouvées deux hypothèses distinctes sur le comportement des non-répondants. L'influence du choix du modèle sur les statistiques produites à partir de l'enquête semble limitée.

Collecte et enquêteurs

Toujours dans le contexte des enquêtes auprès des ménages, Catherine Berthier et Bernard Neros (Insee-DG, unité Méthodes statistiques) ont présenté une méthode de mesure de l'effet enquêteur, défini comme la part de variation des réponses due au fait que tous les ménages ne sont pas interviewés par le même enquêteur. La méthode, qui a été testée sur la première vague du panel européen, consiste à couper en deux les zones géographiques échantillonnées et à les répartir sur des paires d'enquêteurs. On calcule alors, pour une statistique donnée, la différence entre les deux échantillons, et on la situe sur une distribution empirique, construite par mélange aléatoire des réponses des ménages. Une conclusion est que l'effet enquêteur est variable selon les questions posées, mais ne joue pas sur la non-réponse totale (toutes causes de non-réponse confondues).

La plupart des enquêtes statistiques auprès des entreprises sont encore réalisées par courrier. Mais ce mode de collecte est mal adapté au cas des grandes entreprises, dont la structure ne cesse de se complexifier. Il faut également compter avec le rejet

2. Cf. l'article de Dominique ANSIEAU, Chantal CASES et Christine CHAMBAZ : « Le panel communautaire de ménages », *Courrier des statistiques* n° 83-84, décembre 1997.

accru de la charge administrative, qui obère les taux de réponse. De fait, le recours aux enquêteurs apparaît de plus en plus nécessaire. Chantal de Barry (Insee-DG, cellule Coordination des activités d'enquêtes) et Marcel Perrot (Insee-DG, division Harmonisation des enquêtes auprès des entreprises) ont présenté une expérimentation visant à initier un réseau d'enquêteurs terrain, sur des opérations de faible volume. Le premier bilan est positif : amélioration du taux de réponse, de la qualité des réponses et des relations avec les entreprises. Une expérimentation plus poussée est prévue pour 1998, avant le déploiement complet du dispositif.

Enfin, Philippe Houssay (Insee-DG, pôle Infrastructures géographiques) a exposé les grandes lignes du projet CICN, cartographie infracommunale numérisée : cartographie automatique des districts du recensement, cartographie thématique à l'îlot et géocodage à l'adresse. Une démonstration de cartographie numérisée pour les immeubles de la Réunion a été effectuée.

Le logiciel de calcul de précision POULPE

Nathalie Caron (Unité Méthodes statistiques) a présenté les aspects méthodologiques du logiciel POULPE, programme optimal et universel pour la livraison de la précision des enquêtes, qui permet de calculer des statistiques simples (totaux) ou complexes (ratios ou fonctions de plusieurs variables) ainsi que la variance de ces estimations pour des plans de sondages très divers : sondage aléatoire simple, sondage à plusieurs degrés, en plusieurs phases, prise en compte de la non-réponse ou du calage sur marges.

Jean-Noël Petit (Insee-DG, unité Méthodes statistiques) a ensuite exposé les aspects informatiques de ce logiciel, programmé en SAS. POULPE modélise un sondage complexe par un arbre, dans lequel chaque branche renferme les caractéristiques d'un sondage élémentaire : type de tirage, unités tirées, variables auxiliaires, etc. A partir de ce modèle et des fichiers de données, le logiciel applique les formules de calcul (probabilités d'inclusion et variance) et produit les statistiques demandées avec leur précision.

Dans le cadre de l'exploitation de l'enquête logement, David Le Blanc (Insee-DG, division Logement) a notamment estimé au moyen du logiciel POULPE la répartition des ménages par grand statut d'occupation du logement (propriété, location HLM, location libre) et les flux quadriennaux afférents. Il a pu montrer, par le calcul de l'intervalle de confiance, que la progression apparente de la proportion de propriétaires entre 1992 et 1996 était en réalité une stabilisation. Il a souligné le rôle important que peut jouer POULPE pour inciter les statisticiens à calculer systématiquement la précision des estimations et à la communiquer aux utilisateurs.

Indices

Les substitutions que les consommateurs effectuent entre les produits ou entre les lieux d'achat peuvent être prises en compte dans l'indice des prix à la consommation grâce à une technique de chaînage d'indices. Afin de réduire le délai entre deux chaînages (actuellement un an en France), François Magnien et Jacques Pougnard (Insee-DG, division Prix à la Consommation) ont proposé de recourir à l'utilisation des micro-données produites par les sociétés de marketing. Ces données scannées offrent en effet des informations particulièrement précieuses, tant sur les prix que sur les quantités consommées. Une expérience pratique, et concluante, a été menée à partir de données du Panel AC Nielsen relatives au café.

Les services de télécommunication sont en constante évolution, ce qui rend difficile le calcul des purs mouvements de prix. Dans sa communication, Charles Bérubé (Insee, direction régionale des Pays de la Loire) a discuté l'utilisation d'une valeur unitaire pour calculer un indice de prix à la production de ces services. Il a exposé les avantages et les inconvénients d'un indice à valeur unitaire, et les techniques qui peuvent servir au calcul.

François Lequiller (Insee-DG, département des comptes nationaux) a fait le point sur la polémique relative au biais des indices de prix à la consommation. Une commission, dite Boskin, avait en effet annoncé en 1996 que les prix étaient surestimés de 1,1 % par an dans l'indice des prix à la consommation américain. L'Insee a montré qu'en France le biais était de bien moindre ampleur. Mais de grandes lacunes subsistent dans le calcul des indices. L'exploitation des données scannées améliorera la prise en compte des phénomènes de substitution, mais il faudrait également utiliser les régressions hédoniques pour estimer les changements de qualité, et regrouper les services médicaux.

Faire pour apprendre : trois expériences de formation active aux enquêtes

L'Ensai, Ecole nationale de la statistique et de l'analyse de l'information, a la volonté de compléter les acquis théoriques de ses élèves par la réalisation de projets. Pour les élèves de la filière statistique, il s'agit de réaliser une enquête grandeur nature, répondant à un problème posé par un commanditaire extérieur. En 1996-1997 par exemple, le commanditaire était l'unité d'économie et de sociologie rurale de l'Inra de Rennes, et l'enquête portait sur les différentes attitudes face au problème de la « vache folle ». Yves Tillé et Michel Simioni (Ensai) ont décrit les problèmes pédagogiques posés par l'organisation d'un travail collectif de grande dimension à des fins d'apprentissage : distinction entre milieu scolaire et milieu professionnel, difficultés organisationnelles liées au risque d'échec, au degré de mobilisation des

élèves, au travail en équipe... Ils estiment qu'en la matière, le rôle des enseignants recouvre quatre grandes fonctions : ressources et conseil, structuration du travail, arbitrage, et évaluation.

Dans le cadre de leur formation à Libourne, les contrôleurs stagiaires de l'Insee doivent réaliser, sur cinq semaines et demie, une enquête complète proposée par un commanditaire extérieur. Bertrand Roucher (CEFIL) a présenté les objectifs de cette opération : élaborer un système de recueil de données, gérer efficacement une enquête, produire des données normalisées après traitement de l'information, restituer et mettre en forme l'essentiel des informations contenues dans un ensemble de données, rédiger une première analyse et la présenter oralement, organiser et réaliser un travail en groupe. Malgré les écueils rencontrés tels que le choix du partenaire, la détermination du sujet et l'organisation pratique, la première expérience, sans être une réussite parfaite, a montré que les objectifs initiaux pouvaient être atteints. La deuxième expérience est en bonne voie. Il faut espérer que ce type d'opération pourra être renouvelé chaque année.

Benjamin Zanou (Côte d'Ivoire) a décrit le cours de pratique d'enquêtes à l'Ensea, Ecole nationale supérieure de statistique et d'économie appliquée et passé en revue les problèmes que suscite son organisation : recours à du personnel extérieur, encadrement, évaluation des étudiants et perception des enquêtes Ensea par la population. Le point de vue des étudiants est plutôt positif, malgré quelques appréhensions pour aller sur le terrain et des séances de classe parfois houleuses.

Conférence inaugurale

HARMONISATION DES MÉTHODES AU NIVEAU EUROPÉEN : UN PRÉALABLE POUR ASSURER LA COMPARABILITÉ OU UN MYTHE ?

Daniel Defays

Introduction

La qualité est devenue depuis quelques années un label qu'il importe d'afficher ; certains en font un argument de vente, d'autres un mode de gestion. Les services publics ne paraissent pas pouvoir échapper à cette tendance. Plus que d'une mode, il s'agit d'une préoccupation profonde liée à des exigences croissantes des consommateurs, un souci d'efficacité des gouvernants, une pression d'un environnement de plus en plus concurrentiel. Certains offices statistiques ont déjà réagi ; ils proposent à leurs clients, leurs fournisseurs et leurs partenaires des chartes qualité, introduisent des modes de gestion basés sur la qualité totale (TQM), envisagent de se faire certifier. Eurostat, l'office statistique des Communautés européennes, n'a pas échappé à ce mouvement. Il a défini sa mission en termes de fournitures de services statistiques de *qualité.* Cette ambition a replacé la comparabilité, qui au niveau européen est une composante essentielle de la qualité, au centre des préoccupations. Elle fait actuellement l'objet de nombreux débats : ne peut-on pas harmoniser les données a posteriori, faut-il tout harmoniser au même niveau, l'harmonisation complète a-t-elle un sens dans des contextes socio-économiques et surtout administratifs différents ? L'objet de cet article est de faire le point sur ces questions, en insistant particulièrement sur l'impact que différents choix méthodologiques peuvent avoir sur le niveau de comparabilité des données. Il débute par un rappel de ce qui justifie une plus grande harmonisation des données européennes et de ce qu'on appelle comparabilité, il examine ensuite à partir d'études récentes la possibilité réelle d'harmoniser *a posteriori* ou plutôt l'interaction entre mesures et méthodes de mesure. La possibilité de distinguer différents niveaux de comparabilité est ensuite évoquée avant d'aborder une question plus fondamentale, centrale pour la construction européenne : l'harmonisation est-elle possible et si la réponse est oui, que signifie exactement ce terme ?

Pourquoi harmoniser ?

Il est devenu commun de souligner les différences entre données, information, connaissance. Quel que soit le point de vue adopté, il apparaît clairement que la valeur informative d'une donnée est liée à sa capacité à renvoyer à d'autres informations, à ses connotations. Un nombre isolé ne signifie rien. Il faut lui attacher des éléments descriptifs, le situer dans le temps, le comparer à d'autres pour en faire naître une signification. La statistique n'échappe pas à cette logique. Un nombre d'entreprises innovantes dans un pays, un PNB, s'apprécient aussi par référence à des données similaires (comparables ?) collectées dans d'autres pays, à d'autres moments. Obtenir des données comparables apparaît donc dans un premier temps comme une nécessité pour donner de la substance, du contenu à nos informations statistiques (voir par exemple le débat sur ce thème en Intelligence Artificielle, Hofstadter, 1982).

La globalisation, si souvent évoquée, appelle également une internationalisation de nos données et par conséquent plus de comparabilité ; les chefs d'entreprises opèrent sur des marchés qui dépassent les frontières nationales, les pays échangent des biens et des services, les citoyens voyagent. Les pouvoirs publics doivent se préoccuper de ces flux, de ces échanges pour comprendre leur propre économie et décrire la société, les acteurs socio-économiques réclament plus d'informations sur ce qui se passe à l'étranger, un sentiment de citoyenneté supra nationale est en train d'émerger et suscite un intérêt croissant pour ce qui se passe au-delà des périmètres nationaux.

Faut-il mentionner ici la construction européenne, l'apparition d'une administration spécifique avec ses besoins propres en information sur l'Union ? Les politiques communes agricole, régionale, de recherche et développement, la mise en place d'un réel marché intérieur nécessitent à des fins de gestion des informations comparables et souvent qui puissent être agrégées au niveau européen. Ces données conditionnent la gestion de budgets impressionnants. L'utilisation, comme une des références pour définir les contributions nationales au budget communautaire, du PNB, concept central de la comptabilité nationale, elle-même principal élément d'articulation des systèmes statistiques, a également eu une influence déterminante sur l'harmonisation des statistiques.

Et le plaidoyer pourrait continuer. La comparabilité internationale est devenue une exigence incontournable en cette fin de vingtième siècle ; elle correspond à une demande forte des utilisateurs de statistique exprimée à de nombreuses occasions. Elle conditionne par conséquent la qualité de nos services, si cette qualité est définie, conformément à la norme ISO 8402, comme l'ensemble des propriétés et des caractéristiques d'une entité qui lui confèrent l'aptitude à satisfaire des besoins exprimés et implicites.

Si la reconnaissance de la nécessité de plus de comparabilité ne paraît pas poser de problème, la définition de ce qu'on entend par comparabilité est moins évidente.

Que signifie harmonisation ?.

Dans le début de cet article nous avons utilisé de manière quasi interchangeable les mots ‘comparabilité’ et ‘harmonisation’. Leur signification est pourtant différente et mérite d’être précisée. L’harmonisation paraît un préalable à la comparabilité. Ceci est pourtant un peu court. Le recours aux dictionnaires peut aider à mieux comprendre les significations et rôles respectifs de ces deux concepts. Comparer, c’est ‘examiner les rapports de ressemblance et de différence’ ou c’est ‘rapprocher en vue d'assimiler ; mettre en parallèle’, nous dit ‘Le petit Robert’, alors qu’harmoniser, c'est ‘mettre en harmonie, en accord ‘, l'harmonie étant elle-même définie comme les relations existant entre les diverses parties d'un tout qui font que ces parties concourent à un même effet d'ensemble. On peut ainsi comparer des statistiques relativement différentes comme le nombre de chômeurs dans une région donnée à une époque donnée avec la population totale correspondante ou des dépenses de fonctionnement moyennes dans une population d'entreprises avec un chiffre d'affaires moyen ; ces comparaisons ont un sens parce qu’elles permettent de mettre les données en parallèle, d’opérer des assimilations (de calculer des rapports par exemple), mais ces données ne doivent pas nécessairement être harmonisées. L’harmonisation est plus exigeante ; elle requiert l'existence d'un tout, une subdivision en parties, et l’existence d'un accord, d'une correspondance entre ces parties. Le fait d’utiliser de manière indifférenciée les mots comparabilité et harmonisation est donc un abus de langage. Dans ce qui suit, nous nous intéressons à la notion stricte d'harmonisation. Un des objectifs de cet article est du reste de lui donner un contenu plus précis.

La distinction étant établie, il importe cependant de répéter qu’une des justifications de la nécessité d'harmoniser est de pouvoir comparer des données d'origines différentes : statistiques relatives à différents secteurs, différentes périodes, différentes zones géographiques. La comparaison a pour objet de mettre en rapport des différences ou des ressemblances observées avec des effets liés aux secteurs, au temps, aux pays. Pour que ce rapprochement soit possible et surtout instructif, il importe que les effets soient aussi purs que possible. Des écarts qui sont attribuables en même temps à des différences méthodologiques (définitions de concepts différentes, méthodes de mesure non similaires ...) et à la variable explicative d’intérêt sont difficiles à interpréter. Si l’impact de la méthodologie peut être assimilé à un bruit, ceci n’est pas gênant pourvu qu’on puisse estimer les niveaux de précision atteints dans l'estimation des paramètres étudiés. N'est-il pas courant dans les sciences expérimentales de comparer des caractéristiques de population à partir d'estimateurs de variance différents ? Malheureusement la manière précise dont les différents choix méthodologiques opérés affectent les grandeurs estimées est généralement inconnue. En comparant deux pays, on risque donc d'interpréter des biais liés essentiellement à des méthodes différentes. Ceci semble plaider pour une harmonisation complète des méthodes de mesure. Cette position radicale rencontre comme on peut s’en douter de nombreuses résistances.

L'harmonisation *a posteriori*

La comparabilité, comme je l'ai signalé en début d'article, n'est qu'une des composantes de la qualité, à côté de la pertinence de la mesure ('relevance' en anglais), de sa précision, de sa fraîcheur, de sa cohérence avec les autres informations statistiques, par exemple (voir, par exemple, Depoutot, 1998). Dans certains cas, des conflits peuvent apparaître entre ces différentes exigences. Les enquêtes nationales visent généralement à apprécier des évolutions temporelles, des effets régionaux ou sectoriels ; la comparabilité internationale est alors perçue comme un bénéfice complémentaire mais qu'on n'est pas prêt à percevoir à n'importe quel prix. Les systèmes nationaux possèdent leur propre logique, leur propre cohérence, doivent répondre à des exigences de fraîcheur des données qui peuvent entrer en conflit avec les prescriptions internationales. Un concept harmonisé peut perdre sa pertinence dans certains pays. Ceci conduit à des arbitrages. Certains pays ont défendu le concept d'harmonisation a posteriori. Les méthodes de mesure restent de la compétence exclusive des pays et ceux-ci opèrent en fin de traitement les corrections nécessaires pour rendre les données comparables. Des données sur la population seront donc collectées par recensement dans certains pays, par exploitation de registres administratifs dans d'autres, par questionnaire ou par entretien, avec ou sans correction pour les non-réponses. L'idée sous-jacente est qu'il existe un concept, par exemple un paramètre de population précisément défini, qui peut être mesuré indifféremment de différentes manières. Les spécifications communautaires doivent porter sur la nature des résultats désirés et non sur les méthodes à utiliser pour les obtenir. Cette position n'est pas dénuée d'ambiguïté. Où finissent les résultats et où commencent les méthodes ? Admettons que l'on se soit mis d'accord sur ce qu'est l'innovation technologique, peut-on réduire les obligations communautaires au comptage, par exemple, des unités innovantes ? Non, bien sûr ; il faut encore spécifier le type d'unités, la population et l'époque de référence. Peut-on s'arrêter là en décrétant que la statistique qui nous intéresse est l'effectif de l'ensemble des entreprises innovantes du secteur manufacturier à la date du premier janvier 1993 ? Suivant que l'information est recueillie par téléphone ou par entretien individuel, par une interrogation du chef d'entreprise ou d'un comptable, par une question posée à la fin ou au début d'un long questionnaire, suivant que le taux de réponse est de 30, 60 ou 90 %, suivant qu'il y a eu correction ou non pour non-réponses, suivant que l'information est produite par exploitation d'un fichier administratif, par recensement ou sondage, obtiendra-t-on des résultats comparables ? La réponse est clairement non. Pour expliquer l'interaction entre méthode de mesure et phénomène mesuré Gribbin (Gribbin, 1984) reprend une anecdote racontée par le physicien quantique Wheeler. Celui-ci fut invité à jouer pendant un dîner au vieux jeu des devinettes. Il sortit de la pièce afin que les autres invités puissent décider quel était l'objet à découvrir mais fut exclu pendant un temps incroyablement long, situation qui attestait que ses partenaires choisissaient un mot singulièrement difficile ou s'apprêtaient à lui jouer un tour. De retour dans la pièce, il constata que les réponses à ses questions du type 'est-ce que ça vole ?", ou "est-ce un objet inanimé ?" étaient d'abord très rapides, mais au fur et à mesure que le jeu avançait, se faisaient de plus en plus lentes. Ceci lui paraissait étrange puisque le groupe était supposé s'être mis d'accord a priori sur un objet et qu'il suffisait de répondre par oui ou par non. Après un long interrogatoire de l'assemblée , Wheeler suggéra : "est-ce un nuage ?". La salle

répondit "oui" en choeur et dans un éclat de rire. En fait, ses amis s'étaient mis d'accord non sur l'objet qu'il convenait de deviner, mais sur le fait que chaque personne interrogée devait donner une réponse sincère concernant un objet réel auquel elle pensait et qui devait correspondre à toutes les réponses précédentes. Chaque joueur devait donc imaginer son propre objet et le modifier progressivement en fonction des réponses données aux questions de Wheeler. Chacun, à chaque réponse, révisait, le cas échéant, l'objet auquel il pensait, d'où la difficulté de la tâche non seulement pour Wheeler mais également pour ses amis. En opérant ainsi, l'objet qu'a découvert Wheeler est un pur produit du processus utilisé pour le trouver. Il est construit par les questions posées et les réponses données. Il est impossible dans certains cas de dissocier méthode de mesure et objet de mesure. Cet exemple emprunté à un exposé sur la mécanique quantique peut être transposé à la statistique ou du moins à certains domaines de la statistique. Comment définir une attitude par exemple sans faire référence au questionnement ?

L'impact des modes de collecte et de traitement de l'information sur les résultats obtenus est un sujet qui préoccupe Eurostat depuis longtemps. Dans les paragraphes qui suivent je présente brièvement quelques travaux récents sur ce sujet. Ils permettent de mieux apprécier les rapports complexes qui existent entre les instruments de mesure, les méthodes et les mesures réalisées.

L'étude de Statistique Pays-Bas sur l'interprétation des unités statistiques

Faire de la statistique, c'est estimer des paramètres de population. Les populations sont constituées d'unités sur lesquelles il importe de se mettre d'accord si l'on veut comparer des résultats. L'importance de ce problème est du reste reconnue par l'existence d'un règlement communautaire sur les unités statistiques. Les prescriptions internationales ne peuvent être formulées qu'en termes généraux compte tenu de l'hétérogénéité des situations nationales. Comment les pays interprètent-ils ces normes communautaires et quel est l'impact de ces différentes interprétations sur la comparabilité des résultats ? Ces questions ont fait l'objet d'une étude commanditée par Eurostat et exécutée par Statistique Pays-Bas. Pour ce faire, trois pays ont été invités à échanger des informations sur des unités statistiques et à comparer les délinéations obtenues pour ces unités en utilisant leurs propres critères. Plus précisément, chaque pays a identifié dans son répertoire national 16 ensembles d'unités (des secteurs manufacturier et des services) qu'il a proposés aux collègues des deux autres pays. Ceux-ci avaient pour tâche de définir dans ces ensembles ce qu'ils considéraient être des entreprises. Pour ce faire, ils étaient autorisés à poser des questions supplémentaires au pays donateur : existence de liens financiers, situation géographique précise, interdépendance des gestions etc. A partir de ces informations, chaque pays a proposé une subdivision en entreprises, au sens national du terme, des ensembles fournis par les deux autres et bien entendu de l'ensemble que lui-même avait proposé. Trois variables ont été attachées à ces unités : le secteur d'activité, l'emploi et le chiffre d'affaires. Trois délinéations différentes (une pour chaque pays) d'un même ensemble d'opérateurs économiques (environ 45 au total)

ont ainsi été définies. Elles correspondent à trois interprétations d'une même norme communautaire (en fait, 5 interprétations différentes ont été proposées deux pays ayant une interprétation théorique différente de celle qui était effectivement utilisée). Les résultats obtenus sont inquiétants, un pays a dans le secteur des services identifié 21 entreprises là où un autre n'en avait trouvées que 7. L'impact sur les totaux par secteur n'est pas non plus négligeable : des différences de l'ordre de 10% ont été observées dans certains cas. Les résultats de cette étude sont bien entendu difficiles à généraliser, les ensembles initiaux n'étant pas représentatifs des populations généralement étudiées.

Il invite cependant à la prudence et montre au minimum la nécessité de confronter les interprétations nationales pour mieux comprendre les disparités observées.

La population de référence

Un autre exemple intéressant de sensibilité des résultats à différentes interprétations données à un même concept est donné par Sirilli (Sirilli, 1997). Il s'agit cette fois-ci de la définition de la R&D. Dans le manuel de Frascati, le champ des enquêtes sur la recherche et le développement est défini à partir de la notion de recherche systématique ("travaux entrepris de façon systématique"), c'est-à-dire effectuée de manière continue. Cette restriction, comme le montre l'étude de cas qui suit, est fondamentale. Pour l'année 1992, l'Italie lors de deux enquêtes successives, une sur la R&D proprement dite et l'autre sur l'innovation, a utilisé des variantes. Dans l'enquête R&D, le champ couvre l'ensemble des entreprises qui effectuent de la R&D de manière stable et continue et qui ont une certaine taille ; dans l'enquête innovation, sont couvertes les firmes, quelle que soit leur taille, qui ont innové mais sans effectuer nécessairement elles-mêmes de la recherche. Une question commune permet de comparer le nombre de firmes qui font de la R&D et leurs dépenses. Il est normal de s'attendre à des différences puisque les populations de référence sont en fait différentes. Mais leur ampleur surprend : 748 enterprises pour l'enquête R&D et 4229 pour l'enquête innovation. Et des différences similaires sont observables en matière d'évolution de 1985 à 1992 (de 793 à 748 dans un cas, de 2557 à 4229 dans l'autre). Par contre les données de dépenses - comme on pouvait s'y attendre puisque les différences sont sûrement attribuables à l'inclusion des petites enterprises et de recherches occasionnelles dans une enquête - ne diffèrent que de 14%. Cet exemple bien entendu ne démontre rien du tout puisque les champs comme je viens de le préciser étaient différents. Il confirme la sensibilité des statistiques à la définition des populations de référence, ce qui est bien entendu une évidence, et invite à se méfier de l'impact que pourraient avoir différentes opérationalisations de la notion de recherche systématique sur les statistiques produites. De légères variantes risquent d'introduire des variations importantes et d'invalider des comparaisons internationales. Sans opérationalisation commune de la notion de recherche systématique les comparaisons pourraient être vaines....

L'importance du questionnaire

L'importance de la formulation et de la présentation des questions sur les réponses qui leur sont données est connue depuis longtemps. Des travaux récents de Piazza et Sniderman (Piazza, 1997) ont par exemple montré que des écarts considérables – passage d'une proportion d'approbation de 26 à 63 % - sont possibles suivant la manière dont une question est introduite et formulée. L'exemple donné porte sur la discrimination raciale. La question posée concerne la préférence qu'il faut donner en matière d'admission à l'université aux candidats présentant les qualifications nécessaires qui sont noirs de peau. Dans une première version du questionnaire, la question était précédée d'une note explicative qui soulignait l'existence dans le passé d'une discrimination qui avait profité aux blancs. Elle insistait sur la nécessité de compenser cet état de fait en donnant préférence aux noirs ; elle reconnaissait que ce point de vue n'était pas partagé par tout le monde, certains n'admettant pas l'utilisation de critère racial pour la sélection à l'université. L'autre version était assez semblable : une note explicative reconnaissait une discrimination dans le passé et de la nécessité de consentir un effort supplémentaire pour assurer que les noirs qui présentent les qualifications nécessaires soient considérés. Le point de vue opposé (pas de discrimination basée sur la race) était également présenté comme dans la première version. Sur les 889 personnes interrogées à partir de la version 1, 26% se sont déclarés en faveur d'une discrimination positive à l'égard des noirs. Ce pourcentage, calculé sur un échantillon de 911 personnes, est passé à 63% avec la version 2 où la nécessité d'un effort supplémentaire était mentionnée dans la question.

Ce qui étonne n'est pas l'existence d'un effet mais plutôt son ampleur. Et il n'est pas sûr que l'exemple donné soit caricatural. Les mots utilisés dans les questions sont quelquefois entourés de halo, de connotations qui ont souvent des origines culturelles et qui peuvent à eux seuls induire des différences non négligeables. Pensez simplement au concept d'innovation par exemple ou au mot développement qui admet différentes traductions dans certaines langues.

D'autres illustrations de l'influence de la manière dont sont formulées les questions existent. Van Bastelaer, par exemple, cite l'exemple de la statistique de l'emploi aux Pays-Bas où l'accroissement relatif du nombre de femmes qui ont un emploi rémunéré entre 85 et 88 est respectivement de 26 % ou 13 % suivant l'enquête de référence utilisée (enquête force de travail ou enquête auprès d'établissements). Selon l'auteur, cette différence est directement attribuable au mode de questionnement (Van Bastelaer, 1994). Il cite également l'exemple de l'Allemagne qui a modifié le questionnaire de son mini-census en 1990 en ajoutant une question qui faisait référence explicitement à l'exécution de travaux dits mineurs (moins de 15 jours par semaine, salaire inférieur à 450 DM et pas de contribution à la sécurité sociale), alors que ces travaux étaient déjà couverts dans les enquêtes précédentes (des instructions explicites à cet égard avaient été données aux enquêteurs). Suite à cette modification du questionnaire, la proportion de personnes exerçant un travail de cette nature est passée de 2 % en 1988 à 4 % en 1990.

La correction pour non-réponses

De plus en plus d'instituts nationaux de statistiques sont confrontés à des problèmes importants de non-réponses. Dans certains pays, il n'est pas exceptionnel pour des enquêtes non obligatoires d'observer des taux de réponse inférieurs à 50 %. La première enquête communautaire sur l'innovation est particulièrement illustrative à cet égard : 33 % en Irlande, 22 % en Allemagne, 50 % aux Pays-Bas etc. Dans des situations aussi extrêmes un examen attentif des non-répondants s'impose. Ceci peut se faire de différentes manières. Certains se sont contentés des résultats rassurants obtenus en Allemagne où aucun biais attribuable aux non-répondants n'a été mis en évidence et n'ont pas organisé d'enquête complémentaire ; ils ont simplement adapté leurs facteurs de pondération en conséquence. D'autres ont sondé les non-répondants et ont appliqué des corrections. A priori, les profils de non-répondants dans les différents pays, soumis à une même enquête harmonisée, ne devaient pas être substantiellement différents. Les corrections n'auraient pas dû changer les positions relatives des pays en matière de tendance à innover (chiffrée par le pourcentage d'entreprises innovantes). Les résultats ne confirmèrent absolument pas cette intuition. En Irlande par exemple, avant correction, le taux d'entreprises innovantes était de 71,2 % ; après correction il est devenu 33% alors qu'aux Pays-Bas, il est passé de 54,4 % à 58,4 % .

De nouveau ces résultats semblent confirmer des évidences. Ils montrent le soin qu'il convient d'apporter à des recommandations internationales si l'on veut minimiser les disparités liées aux méthodes de mesure et de traitement. Faut-il toujours imposer une enquête auprès des non-répondants ? Si non, à partir de quel seuil est-elle nécessaire ? Sous quelle forme (interrogation téléphonique, envoi d'un questionnaire simplifié, visite...) ? Il n'est pas certain que si ces choix sont laissés aux pays, ils ne soient pas source de biais comme dans l'exemple donné ci-dessus.

Les facteurs de pondération

Très souvent, l'objectif des enquêtes statistiques est d'estimer des totaux de variables sur des populations données (emploi total, somme des valeurs ajoutées, etc.). Pour ce faire, il est courant d'utiliser des estimateurs de type Horvitz-Thompson, où les observations de l'échantillon sont pondérées par les inverses des probabilités de tirage. Souvent, cependant, il est possible de prendre en compte des données auxiliaires (généralement les marges connues de certains tableaux à construire), et d'intégrer ces informations dans l'estimation du total. Cette pratique est courante dans les Etats membres et généralement non pilotée par des recommandations communautaires : quel type de données auxiliaires faut-il prendre en compte, comment "caler" les estimations à partir des données auxiliaires ? Ici encore, les différents choix possibles risquent d'induire des biais et par conséquent d'affecter, si ces biais ne sont pas mesurés ou corrigés, la comparabilité des résultats. Eurostat, dans le cadre du traitement d'une enquête sur l'impact du marché intérieur sur les entreprises, a réalisé certaines simulations qui peuvent aider à mieux comprendre l'ampleur de l'impact de ces différents choix méthodologiques. Un institut

allemand a, en effet, envoyé dans le cadre d'une enquête conduite en 1994 auprès de 1268 entreprises, des données individuelles, des facteurs de pondération et des données agrégées. Les poids fournis ont été recalculés suivant différentes méthodes :

- post-stratification à partir de données auxiliaires sur la distribution des entreprises allemandes par secteur et classe de taille (croisements) ;

- calage sur les marges du tableau précédent (c'est-à-dire totaux par secteurs et totaux par classe de taille) au moyen de différentes méthodes de minimisation de la distance entre les poids initiaux fournis par l'institut et de nouveaux poids (les distances sont calculées par des moindres carrés généralisés - MCG - par itération du quotient - IQ, par une méthode MCG restrictive, ou par une méthode IQ restrictive). Le lecteur intéressé trouvera dans (Bienvenue, 1998) plus de détails sur ces simulations.

Les résultats montrent que si le choix de la distance ne paraît pas affecter fortement les estimations des totaux (on passe, par exemple, de 41 % de "D'accord avec le fait que le programme du marché unique a été un succès pour mon entreprise " à 40,5 %), l'utilisation non seulement des marges, mais également des interactions entre secteur et taille pour post-stratifier les données a une influence considérable, certaines estimations passant de 48 % à 27 % par exemple.

Différents niveaux de comparabilité

Ces différents exemples semblent indiquer que la comparabilité est une exigence fort coûteuse. Soit les méthodes de mesure sont harmonisées, c'est-à-dire dans ce cas unifiées (et ce mot possède des connotations de pensée unique qui ne plaisent pas), et les résultats peuvent être mis en parallèle, les différences peuvent être interprétées, soit les pays sont libres à l'intérieur de recommandations communautaires d'effectuer les choix les plus indiqués pour leur système national, et la comparabilité des données devient faible. Ce dilemme a amené certains à préconiser une harmonisation à différentes vitesses ou à géométrie variable pour reprendre une expression du jargon communautaire (voir par exemple, l'évaluation de l'enquête communautaire sur l'innovation). Le niveau de comparabilité ne doit pas être identique pour toutes les statistiques. Elle a un prix que l'on n'est prêt à payer que dans des circonstances particulières, lorsque par exemple, une politique communautaire l'exige, comme pour la mesure du PNB, l'indice des prix à la consommation ou le chômage. D'un point de vue statistique, la notion de différents niveaux de comparabilité peut s'aborder autrement.

En fait la possibilité même de comparer des données renvoie à la théorie de la mesure ; elle établit quand il paraît justifié pour une mesure donnée (disons une mesure nominale comme le numéro d'une carte d'identité) de comparer, voire de combiner par des opérations arithmétiques données, des valeurs distinctes. On peut ainsi comparer des numéros d'identité et établir s'ils sont identiques ou distincts mais on ne peut pas les ordonner. Pourquoi ? Parce que mesurer c'est établir une correspondance entre un

système formel et un système empirique et que les seules opérations formelles permises sont celles qui correspondent à des opérations identifiées et sensées du système empirique : on peut comparer deux poids de manière empirique avec une balance par exemple mais on ne peut pas faire de même avec des identités : le poids est une mesure ordinale (au moins), l'identité est nominale. Malheureusement la théorie de la mesure se préoccupe assez peu à ma connaissance du problème qui nous intéresse ici, à savoir la possibilité de comparer des données issues de populations différentes.

Nous pouvons pourtant reprendre certains principes de cette théorie pour formaliser la notion de comparabilité et de niveaux de comparabilité.

Supposons par exemple que nous voulions étudier les dépenses moyennes consacrées à la recherche et au développement dans deux zones géographiques distinctes. Il apparaît raisonnable d'exiger pour que les deux moyennes puissent être comparées que

- les valeurs individuelles puissent être comparées ;

- les moyennes puissent être calculées.

En termes de théorie de la mesure, ceci signifie que les deux mesures sur les deux populations comparées doivent être des mesures d'intervalle, sinon le calcul d'une moyenne n'aurait pas de sens. Mais ceci n'est pas suffisant, pour comparer les moyennes ; il faut encore que, si a est un objet quelconque de la première population (une entreprise de la première zone géographique) et si b appartient à la seconde, l'inégalité $f(a) > g(b)$ ou $f(a) < g(b)$ - en notant respectivement f et g les mesures des dépenses de R&D dans les deux zones - ait un sens. En d'autres mots, il faut qu'il existe une relation W entre les deux populations qui soit telle que aWb si et seulement si $f(a) > g(b)$. Dans notre exemple, ceci signifie que l'on puisse comparer empiriquement les dépenses de deux entreprises de zones différentes. Remarquer que cette exigence est moins triviale qu'on ne pourrait croire car si les zones correspondent à des pays différents les dépenses sont en principe libellées en monnaies différentes et ne sont donc pas comparables a priori. De plus, si les niveaux de vie sont très différents, si les coûts d'équipement et les salaires sont beaucoup plus élevés dans une population que dans l'autre, la définition de la relation W doit être faite avec soin.

Une approche complémentaire du sens d'une comparaison de nos deux moyennes est également possible. Elle nécessite l'introduction de quelques notations. Soit A la population d'entreprises dans la première zone géographique et B la population correspondante de la deuxième zone. Les effectifs de ces populations sont respectivement N(A) et N(B). La question que nous nous posons est : quand l'affirmation $(S\ f(a)/ N(A)) > (S\ g(b) / N(B))$ a-t-elle un sens ? La théorie de la mesure ne répond pas directement à cette question. On peut cependant déduire facilement de ses principes mêmes que la comparaison a un sens seulement si (la condition est nécessaire mais sûrement pas suffisante) pour toutes transformations admissibles f de f et g de g, l'inégalité ci dessus est toujours respectée :

$$(S\ (f \circ f)(a)/ N(A)) > (S\ (g \circ g)(b) / N(B))$$

la classe des transformations admissibles étant définie par la nature des échelles f et g. Ceci signifie entre autres que si f et g sont libellés dans une monnaie commune (en écus par exemple), le passage à une autre monnaie commune ne devrait pas affecter le sens de la comparaison.

L'exigence de comparaison ordinale impose donc des contraintes aux mesures f et g. Mais d'autres exigences sont envisageables. Ainsi, il existe souvent un besoin fort de pouvoir constituer à partir de données relatives à des parties des totaux, des agrégations. Pour comparer les taux de chômage en Europe et aux États-Unis il faut combiner des données nationales pour pouvoir obtenir un total européen. La théorie de la mesure spécifie des conditions nécessaires et suffisantes pour que des mesures relatives à différents individus puissent être sommées, multipliées, combinées, à l'intérieur d'une même population, malheureusement, comme je l'ai déjà signalé, sans vraiment se préoccuper des problèmes que peuvent poser des mesures définies sur des populations différentes.

Pour calculer un total communautaire, il faut bien entendu pouvoir sommer des valeurs au sein d'un même pays, ce qui exige une mesure d'intervalle, c'est-à-dire, entre autres, une relation binaire o telle que $f(a \circ b) = f(a) + f(b)$ en reprenant les notations introduites précédemment. Mais ceci n'est pas suffisant. Il faut encore pouvoir additionner des mesures sur des populations différentes. Ceci nécessite d'étendre des opérations définies séparément sur des ensembles A et B à A U B. On devra ainsi s'interroger sur le sens qu'il y a à additionner des dépenses d'entreprises dans des pays différents, c'est-à-dire soumises à des fiscalités différentes et libellées initialement dans des monnaies différentes. De plus, il importe que l'opération qui associe des unités de A et de B soit de même nature que celle qui associe des unités de A ou des unités de B. Je m'explique. Si je compte par exemple des unités innovantes dans un pays A et que je les ajoute à celles d'un autre pays B, implicitement j'assume qu'il est neutre de supprimer une unité dans A et de l'ajouter dans B. Une entreprise innovante italienne est considérée en quelque sorte équivalente à une entreprise innovante belge. Nous retrouvons ici la notion d'équivalence qui soutend tous les propos de cet article, comme je l'explique dans le paragraphe suivant.

Ces considérations pourraient mener à définir différents niveaux de comparabilité basés sur les différents types de traitement que l'on veut faire subir aux données (simples comparaisons, agrégations). Cette approche fonderait théoriquement un concept dont on conçoit l'utilité, voire la pertinence, sans en apprécier les implications précises.

Si les considérations précédentes ont amené certains à proposer différents degrés d'harmonisation, elles invitent également à repenser le concept même d'harmonisation. La question n'est plus de savoir si le coût d'une méthode de mesure unique n'est pas excessif mais de s'interroger sur son sens.

L'harmonisation est-elle possible ?

J'ai abordé ce débat dans un article précédent et j'en reprends ici les éléments principaux (Defays, 1995). Une uniformisation totale des concepts et des méthodes présuppose des environnements géographiques, administratifs, légaux, économiques... identiques. Pour bien faire comprendre cet argument, utilisons des conditions extrêmes. Comparons par exemple la structure des entreprises dans un pays européen et dans un pays en voie de développement. D'un côté des unités sont répertoriées dès qu'elles exercent une activité économique, la main-d'oeuvre est enregistrée, les travailleurs ont des salaires, les systèmes postaux et téléphoniques fonctionnent plus ou moins et les unités sont habituées à remplir des formulaires, des questionnaires. De l'autre les unités sont plus instables, ne sont pas toutes déclarées, l'emploi est peut-être partiellement nomade, le troc existe toujours et les moyens de communication sont plus lents et en général moins efficaces ; l'interrogation systématique par questionnaire n'est pas possible. Que pourrait signifier dans un pareil contexte des méthodes identiques ? Ceci n'a pas de sens et on sent bien dans cet exemple que ce qui est recherché dans une harmonisation ce n'est pas une identité de concepts et de méthodes, mais une équivalence, une correspondance, un accord. Quel type de correspondance ? C'est ce que nous discutons dans le paragraphe suivant.

Quel accord entre les parties du tout pour qu'il y ait harmonisation ?

Le caractère exotique de l'exemple ci-dessus pourrait pour certains en diminuer la pertinence dans les problèmes que nous traitons dans cet article. Les pays de l'Union européenne ne sont pas si différents que cela, tout compte fait. Les mêmes concepts existent plus ou moins ou sont du moins facilement transposables. Toute personne qui fut un jour responsable de l'harmonisation d'un concept au niveau européen vous convaincra facilement du contraire. Prenons une notion aussi centrale et aussi simple a priori que la notion d'entreprise. Supposons que nous intéressions plus particulièrement aux naissances de nouvelles entreprises. Combien de créations en 1997 ? Si la définition d'une entreprise peut paraître difficile, particulièrement parce qu'il en existe de grandes avec des structures complexes, la question à laquelle nous cherchons à répondre ne concerne que les petites entreprises et paraît donc pouvoir s'aborder avec une définition simple et universelle de l'entreprise : une unité légale active. Pourtant un examen de ce concept dans les différents pays fait directement apparaître des différences notables : dans certains pays il n'est pas nécessaire d'être enregistré pour démarrer une activité commerciale : on peut en Angleterre, par exemple, sans aucune licence, ouvrir un petit magasin où l'on vend des fleurs ; de plus dans certains pays les procédures d'enregistrement sont plus longues, plus compliquées et beaucoup plus coûteuses que dans d'autres. La signification même de ce que représente une unité légale en terme de germe d'un futur opérateur économique risque d'en être affectée. Sans mentionner le fait que tous les statisticiens n'ont pas accès aux fichiers d'unités légales et risquent de devoir approcher ce concept au moyen de l'enregistrement d'unité administrative comme l'unité TVA, changeant implicitement la définition du concept. La vraie question à se poser

lorsqu'on veut comparer le nombre de créations d'unités TVA au Royaume-Uni avec ce qui se passe en France par exemple est : "quel est l'équivalent de l'unité TVA britannique en France ?". Le problème est donc un problème d'analogie : l'unité TVA est au Royaume-Uni ce que l'unité X est à la France. Le raisonnement par analogie a été largement étudié en psychologie et plus particulièrement en intelligence artificielle. La statistique pourrait, je crois, utilement s'inspirer de certains de ces travaux. Donnons un nouvel exemple pour mettre en évidence les particularités et les difficultés de ce type de raisonnement (voir D. Hofstdadter, 1995). La première dame des États-Unis est madame Clinton. Qui est la première dame de France ? Madame Chirac ? D'accord mais madame Jospin pourrait être un candidat plausible compte tenu du rôle important joué par un premier ministre en France. Et la première dame du Royaume-Uni ? Ici les choses se compliquent. Le prince Philippe est sûrement candidat mais madame Blair aussi. En effet, la reine d'Angleterre n'est pas un président et le rôle politique de M. Blair est sûrement plus important et plus proche de celui d'un président. Mais dans la notion de première dame, il y a les notions de "premier" et de "dame" et à cet égard la reine Élisabeth paraît un candidat plus approprié. On sent bien dans cet exemple que lorsque les deux structures à comparer ne sont pas identiques il importe de privilégier certaines caractéristiques de la situation initiale du type "être le conjoint du chef d'État ", ou "être très proche du pouvoir", au détriment d'autres aspects du type "être une femme" ou "être le conjoint du président" ; mais ce choix est partiellement arbitraire et dépend des inflexions que l'on veut donner au concept sous-jacent. Des problèmes similaires se posent en statistique. Comment harmoniser une date d'enquête par exemple ? Peut-on considérer que les dates de deux enquêtes - disons agricoles - organisées respectivement en Europe et en Australie sont harmonisées parce qu'elles ont lieu simultanément le 1er avril sur les deux continents ? En termes de saisons, de comportements des agriculteurs, de vacances, cette même date a des implications ou des significations différentes ; harmoniser dans ce cas consisterait sûrement à effectuer les enquêtes à des dates différentes dans le temps mais à une saison identique ou à un moment identique de l'année comptable. Comme dans l'exemple de la première dame des États-Unis, l'analogie nécessite de sacrifier certaines caractéristiques de la situation (sa date dans l'absolu) par rapport à d'autres (date relative par rapport aux saisons...). Harmoniser c'est établir des analogies. Un autre exemple emprunté cette fois à la statistique d'entreprises. Un pays X limite son enquête annuelle sur les entreprises à celles qui occupent 20 personnes ou plus. Ces entreprises représentent 90% de la valeur ajoutée du secteur manufacturier ; elles ont en moyenne un chiffre d'affaires supérieur à 2 millions de piastres (la monnaie nationale). Comment transposer ces concepts dans le pays Y dont les entreprises sont en moyenne beaucoup plus petites ? Une interprétation littérale amène à ne couvrir que les entreprises de 20 personnes ou plus. Malheureusement elles ne représentent que 80 % de la valeur ajoutée totale et leur chiffre d'affaires est en moyenne plus bas. Les deux populations sont-elles comparables ? Ne devrait-on pas prendre les entreprises qui couvrent également 90% de la valeur ajoutée totale en baissant le seuil de taille ? Nous sommes de nouveau confrontés aux choix difficiles du raisonnement par analogie : privilégier certaines caractéristiques au détriment des autres. Mais comment guider ce choix ? Sûrement en fonction des utilisations envisagées des statistiques nationales dans la mesure où elles peuvent être anticipées.

Conclusion

Nous sommes condamnés à comparer des mesures obtenues via des méthodes différentes sur des concepts au mieux équivalents. Mais la nécessité d'identifier et de quantifier des effets nationaux subsiste. Comment résoudre ce conflit ? Différentes voies de solution sont envisageables. Une meilleure compréhension des sources de différences paraît indispensable ; des études du type de celle réalisée par Statistique Pays-Bas devraient probablement être lancées de manière plus sytématique pour mieux contraster les différents choix nationaux et mesurer leur impact sur la comparabilité des résultats. Une théorie de l'erreur qui permette de chiffrer non seulement ce qui est lié aux aléas des échantillonnages mais également aux écarts observés par rapport à des prescriptions communes paraît également nécessaire. Enfin, l'utilisation de modèles dans l'analyse des résultats, comme proposé par exemple par Depoutot (Depoutot, 1997), pourrait permettre de purifier les mesures et d'obtenir des concepts, mesurés par des variables latentes, qui soient plus comparables.

Eurostat entend poursuivre et promouvoir des travaux dans ces différenes directions et invite tous ceux qui sont intéressés par le sujet à joindre leurs efforts au sien.

BIBLIOGRAPHIE

ARCHIBUGI D., COHENDET P., KRISTENSEN A., SCHÄFFER K.A., *Evaluation of the Community Innovation Survey (CIS) - Phase I,* EIMS Publication n°11, Commission européenne, 1994.

BIENVENUE J.Y., "Précision des estimateurs de calage", *rapport interne*, Eurostat, 1998.

DEFAYS D., "Is Harmonization possible ?", *Proceedings of the 1st International Conference on Methodological Issues in Official Statistics in Stockolm*, Statistics Sweden, 1995.

DEPOUTOT R., ARONDEL PH., "International Comparability and Quality of Statistics", W*orking Paper*, à paraître, Eurostat, 1998.

GRIBBIN J., *Le chat de Schrödinger*, Flammarion, 1984.

HERCZOG A., VAN HOOFF H., WILLEBOORDSE A., "The impact of Diverging Interpretation of the Enterprise Concept on Statistical Data", *internal report for Eurostat*, Voorburg, 1997.

HOFSTADTER D.R., *Gödel, Escher, Bach : les brins d'une guirlande éternelle*, Inter Editions, Paris, 1985.

HOFSTADTER D.R., *Fluid concepts and creative analogies*, Basic Books, New York, 1995.

PIAZZA Th., "New Methodological Possibilities Offered by Computer, Assisted Interviewing", *Bulletin of the International Statistical Institute*, Book 1, Istanbul, 1997.

SIRILLI G, "Old and New Paradigms in the Measurement of R&D", *Actes du séminaire CEIES "Statistics on Research and Development", Aarhus 1997*, Eurostat, 1997.

VAN BASTELAER A., "Differences in the Measurement of Employment in the Labour Force Surveys in the European Community", *Journal of Official Statistics*, Vol. 10, N° 3, Statistics Sweden, 1994.

Conférences spéciales

ÉCHANTILLONNAGE DES NON-RÉPONDANTS ET AUTRES MÉTHODOLOGIES POUR LE RECENSEMENT DES ÉTATS-UNIS À L'AN 2000

Yves Thibaudeau

1. Introduction

La conduite du recensement décennal de la population aux États-Unis fait partie des obligations constitutionnelles des branches exécutive et législative du gouvernement. Cette obligation s'insère dans un mécanisme légal qui donne lieu, à tous les dix ans, à une redistribution des districts congressionnaux (équivalents des districts électoraux) entre les états de la fédération. Cette redistribution a pour but de donner aux citoyens une participation au pouvoir aussi démocratique que possible. Idéalement, les rapports entre les nombres de districts congressionnaux de deux états sont fidèles aux rapports de tailles entre les populations de ces mêmes états. Le recensement fournit donc les jalons de la redistribution du pouvoir législatif. Il est impératif que le recensement soit sans biais évident ou erreur grossière. En plus, la nation a le droit d'exiger un recensement de qualité qui utilise des méthodes qui sont à la fois efficaces et économiques.

Le dernier recensement décennal de la population (1990) a entraîné des dépenses de près de trois milliards de dollars. En dépit de ce coût élevé, les biais ainsi que d'autres erreurs de mesure ont empiré par rapport à ceux du recensement antérieur. L'avenir semble promettre une conjecture encore plus difficile pour le recensement de l'an 2000.

Cette situation a entraîné une révision profonde des méthodes et des pratiques suivies dans la conduite du recensement décennal. Rappelons-nous qu'à l'époque du dernier recensement, les organes gouvernementaux au niveau fédéral étaient aux prises avec des problèmes de finances déficitaires qui prenaient alors des dimensions politiques. Le vérificateur général a donc procédé à une enquête diligente dans le but d'identifier les fautes et de suggérer des lignes de conduite qui auraient pour effet de mieux contrôler le coût les opérations de recensement, tout en assurant des résultats fiables. Par ailleurs, le congrès des Etats-Unis décida de parrainer trois commissions d'études, sous l'égide de l'Académie Nationale des Sciences, chacune regroupant plusieurs experts nationaux et internationaux pour en arriver à des

recommandations spécifiques pour corriger les biais et contenir les dépenses. Les recommandations d'une de ces commissions sont publiées par Steffey et Bradburn (1994).

Parmi les recommandations, la proposition d'un échantillonnage des non-répondants est constante. Le contexte dominant est celui d'une économie d'argent importante. Les premières estimations du vérificateur général sont de l'ordre de quatre cents millions de dollars. Ce montant a été révisé à la hausse à cause de divers facteurs, entre autres la situation défavorable à laquelle nous devons faire face sur le marché de l'emploi. En effet, nous prévoyons une compétition vive avec les employeurs qui puisent dans le réservoir de main d'oeuvre où nous nous attendons à trouver nos énumérateurs.

2. Aperçu général de notre stratégie de recensement pour l'an 2000

Le développement logistique du recensement de l'an 2000 s'organise sous quatre thèmes majeurs :

1. Partenariat : Identifier et enrôler des partenaires locaux qui nous aiderons à réussir le recensement.

2. Simplicité : Maintenir une simplicité dans nos opérations, individuellement, et dans l'ensemble.

3. Technologie Intelligente : Acquérir des technologies qui ont des bénéfices immédiats : capture optique digitale, logiciels de couplage.

4. Méthodologie Statistique : Utiliser les méthodologies statistiques pour alléger les opérations, réaliser des économies de fonds, et réduire les biais.

Pour le reste de cette partie, nous révisons les trois premiers thèmes et la façon dont nous les approchons dans nos opérations. Le thème de la méthodologie statistique fait l'objet d'une partie entière.

2.1 Partenariat et marketing

Depuis quelques années nous travaillons à convaincre les institutions locales que c'est dans leur intérêt de coopérer avec nous à la préparation et l'administration d'un recensement de meilleure qualité. Ces nouveaux partenaires sont souvent les gouvernements locaux ou tribaux. Pour le moins ces institutions sont touchées par le recensement, dans la mesure où les octrois de fonds du gouvernement fédéral leurs sont accordés en partie sur la base de la taille de leurs populations. Nous nous

attendons à ce que ces partenaires alertent leurs populations sur l'importance d'une réponse postale au recensement. Nous avons lancé le thème populaire 'Soyez Compté' que, nous l'espérons, les institutions locales vont aider à promouvoir.

Nous avons aussi investi nos énergies à former une alliance avec le service des postes. Celui-ci nous fournira des informations stratégiques d'une valeur inestimable, puisqu'il nous pourvoira des listes d'adresses et des indications quant à l'état de vacance des habitations de chaque quartier. Pour chaque adresse, nous saurons si l'habitation est couramment occupée. Les cas de d'adresses irrésolues sont aussi identifiés.

Finalement, nous avons retenu les services d'une société de marketing privée (Young and Rubicam), dont le contrat est de conduire une campagne de promotion pour la réponse postale au recensement. La diffusion de cette campagne est faite par les médias locaux (télévision, radio, journaux, etc.).

2.2 Simplicité et technologie intelligente

Le thème de la simplicité se traduit par la philosophie de notre approche, plutôt que par une ligne d action définie. Il en va de même pour le concept d'une technologie intelligente.

Un exemple du souci de simplicité se retrouve dans le questionnaire du recensement. De douze en 1990, le nombre de questions est tombé à sept. C'est le strict minimum requis par la loi. Les questions se rapportent au mode de résidence (propriétaire ou locataire), au sexe, à la race, à l'origine hispanique, et à l'âge. Le questionnaire de papier lui-même a fait l'objet d'un nouveau design, visant à faciliter la navigation d'une question à l'autre, réduisant ainsi les chances d'omission accidentelle de certaines questions.

Par technologie intelligente nous entendons une technologie avancée, mais qui en même temps se doit être strictement fonctionnelle. Par exemple, la capture optique digitale de l'écriture sur les questionnaires de papier est une technologie intelligente. Cette technologie nous permet d'abandonner une partie de l'administration onéreuse des questionnaires de papier. Nous estimons que nous encourrons des bénéfices immédiats substantiels en passant à cette technologie nouvelle.

3. Méthodologie statistique

Le Census Bureau a une tradition de conduire des opérations de contrôle de la qualité sur les recensements décennaux. Ces opérations sont déployées selon des plans d'échantillonnage et des procédures statistiques élaborées. Néanmoins, jusqu'à maintenant les analyses de la qualité ne sont que des fins en elles-mêmes. C'est-à-dire que nous faisons toujours part au public du résultat de nos évaluations, mais nous ne nous servons pas de ces mêmes évaluations pour nous guider dans une correction des biais ainsi mis à découvert.

L'argument classique pour éviter une intervention corrective est que les méthodologies statistiques ne sont pas assez fiables pour garantir qu'une correction d'un biais, comme celui du sous-comptage par exemple, n'introduise pas un nouveau biais, encore plus débilitant que le premier. Jusqu'aux années 70 il est certain que cela tenait du vrai. Par contre, aujourd'hui les méthodologistes, presque unanimement, supportent une intervention technique pour amoindrir certains biais. Le biais du sous-comptage et celui de la non-réponse émergent comme des cibles évidentes. Ce dernier est par ailleurs lié aux questions financières, ce qui ajoute à sa pertinence. Nous avons décidé de mettre sur pied deux opérations d'envergure dans le but de corriger ces deux biais et d'alléger les dépenses.

Pour le reste de cette partie nous allons d'abord revoir de façon sommaire la méthodologie de correction du biais du sous-comptage que nous allons utiliser à l'an 2000. Il s'agit d'une méthodologie de capture-recapture à partir d'un plan d'échantillonnage qui chevauche le recensement. Ces techniques ont subi l'épreuve du temps et ont fait l'objet de nombreuses discussions. Ensuite, nous consacrons le reste de l'article à élaborer notre stratégie d'échantillonnage des non-répondants qui permettra une correction du biais de la non-réponse.

3.1 Corriger le biais du sous-comptage

Pour corriger le biais du sous-comptage il s'agit d'étendre la même méthodologie de contrôle de la qualité que nous avons utilisée dans le passé, mais cette fois-ci nous allons intégrer les estimations du sous-comptage dans les statistiques officielles de la population américaine. L'intégration de nos estimations prennent la forme d'ajustements des comptes de la population. Un des problèmes auxquels nous nous sommes heurtés dans le passé est une conséquence directe de la petite taille de l'échantillon que nous avions alors.

En effet, lors du dernier recensement, pour assurer la stabilité de notre inférence, nous avons du diviser notre échantillon en strates définies par le degré d'urbanisation des régions qui en faisaient partie. Étant donnée la petite taille de l'échantillon, et pour obtenir une précision que nous jugeons satisfaisante, il fut

nécessaire de regrouper ensemble des régions provenant de plusieurs états distincts (dans la fédération américaine). Le statisticien comprend bien la valeur de créer des strates homogènes du point de vue des variables qu'on doit mesurer. Pour nous, une dissociation géographique ne pose pas de problème. Malheureusement les gouvernements locaux regardent d'un mauvais oeil une pratique qui a recours à des prélèvements démographiques dans des états éloignés des leurs pour ajuster les comptes de leurs populations.

Pour l'an 2000 nous construisons un échantillon de recapture de 750,000 unités d'habitation. Ces unités sont recueillies en 'blocs' qui sont en fait les unités d'échantillonnage. C'est à dire qu'un bloc entier est sélectionné dans l'échantillon. Un bloc, dans une région urbaine, correspond à un pâté de maisons. Dans une région rurale, on essaye de suivre la même définition, de façon plus ou moins arbitraire. Ces blocs sont canevassés méticuleusement, immédiatement après le recensement. En utilisant des techniques de couplage d'enregistrement, nous identifions les citoyens vivant dans ces blocs qui sont recapturés dans l'échantillon. Nous identifions aussi ceux qui sont capturés une fois seulement, soit par le recensement ou soit par la recapture. Ces trois comptes servent ensemble à produire une estimation du biais du sous-comptage. C'est la méthode de l'estimateur duel. Mulry et Spencer (1992) présentent une excellente discussion de cette technique.

Nous soulignons qu'au prochain recensement la taille de l'échantillon de recapture est telle que nous pouvons estimer le biais du sous-comptage des grandes régions métropolitaines avec une précision suffisante, et ce sans avoir recours à une stratification exogène.

3.2 Corriger le biais de la non-réponse

Nous abordons ici un terrain qui n'a pas encore été défriché. À notre connaissance, aucun recensement national n'a eu recours à un plan d'échantillonnage pour corriger les erreurs causées par la non-réponse. Nous avons déjà mentionné que nous sommes poussés vers cette alternative par nécessité. Au dernier recensement, les unités d'habitation qui n'ont pas participé au recensement postal se chiffrent à plus de trente millions, soit plus du quart des unités. Les opérations de récupérage des non-répondants sont devenues de plus en plus difficiles à administrer. Les énumérateurs tendent à éviter les quartiers à haut taux de non-réponse, ce qui conduit à une erreur difficile à mesurer. Recruter des énumérateurs efficaces est plus difficile que jamais. La situation à laquelle nous faisons face est un marché de l'emploi clairement défavorable. La rémunération que nous offrons aux énumérateurs est compétitive, mais les termes le sont moins : il est clair au départ que le travail d'énumérateur est pour une période de temps limitée. Par contre, l'énumérateur prospectif peut souvent choisir parmi d'autres emplois plus ou moins précaires, dont certains offrent au moins une possibilité de travail à long terme.

Nous avons décidé de rediriger nos efforts à énumérer seulement qu'une portion des non-répondants. Cette portion doit constituer un échantillon aléatoire et nous devons faire en sorte que la qualité de cette énumération réduite excède celle de n'importe quelle énumération des non-répondants qui se voudrait complète. Nous remplaçons donc une erreur opérationnelle, difficile à mesurer, par une erreur d'échantillonnage, que nous savons bien quantifier.

Notre plan d'échantillonnage ne peut être statique, puisque la population des non-répondants n'est connue qu'après le recensement postal. Nous avons donc arrêté une stratégie bien définie. Dans ce qui suit nous définissons notre projet d'échantillonnage. L'unité géographique cible est la 'tract'. Une tract est un voisinage, elle représente en moyenne 1700 unités d'habitation, soit environ 4000 personnes, mais cette taille peut varier considérablement.

La population des non-répondants d'une tract dont le taux de réponse au recensement postal est moins de 85 % est échantillonnée de la façon suivante :

- l'échantillon est aléatoire,

- la taille de l'échantillon est telle que les répondants au recensement postal, ajoutés aux non-répondants échantillonnés, comptent pour au moins 90% des unités d'habitation de la tract.

La population des non-répondants d'une tract, dont le taux de réponse au recensement postal est égal à, ou est plus élevé que 85%, est échantillonnée de la façon suivante :

- l'échantillon est aléatoire,

- l'échantillon couvre le tiers de la population des non-répondants.

Notre but est de définir une stratégie qui s'explique bien dans un langage que nous voulons accessible au citoyen peu familier avec la statistique. Nous pouvons traduire notre plan de façon simple :

Nous garantissons une énumération avec une erreur réduite relativement aux recensements précédents. Nous assurons une énumération directe d'au moins 90% de la population de chaque voisinage.

Cette stratégie n'est pas parfaite du point de vue technique. Par exemple, puisque les tracts sont de tailles différentes et que le taux de la réponse postale varie lui aussi, nous ne pouvons nous attendre à une erreur d'échantillonnage uniforme. Nous décidons d'abandonner des critères scientifiques plus rigoureux, en espérant qu'une bonne compréhension de notre projet amènera une participation accrue.

4. Résultats et conclusion

Farber et Navarro (1997) rapportent les résultats de simulations de scénarios semblables à celui que nous proposons ici. Leurs résultats sont basés sur le recensement de 1990. Les auteurs simulent des stratégies qui, comme la présente, reposent sur le prélèvement d'un échantillon de recapture ainsi que d'un échantillon de non-répondants, quasi-simultanémemt. Les auteurs estiment que l'erreur d'échantillonnage totale au niveau de l'état varie entre .2% et .5%, avec une absence théorique de biais. Ces chiffres se comparent avantageusement avec les erreurs encourues au recensement de 1990 : un biais de sous-comptage moyen estimé à 1.6%, et un biais de non-réponse indéterminé, mais non négligeable. Nous rappelons que ce sont précisément les comptes du recensement au niveau des états qui sont à la base de la redistribution des districts congressionnaux. Le nombre total de districts congressionnaux reste constant. La distribution des districts entre les états seule peut changer, suivant la publication des nouveaux comptes de la population. Une marge de quelques points de pourcentage en plus ou en moins peut signifier un gain ou une perte d'un district. Il est donc très naturel que les milieux politiques s'intéressent à notre stratégie qui utilise des méthodes traditionnelles, aussi bien que des techniques nouvelles. Notre vœu le plus sincère est que cette nouvelle attention portée sur notre méthodologie suscite une participation au recensement plus élevée de la part du public.

Bibliographie

FARBER, J. et NAVARO, A.(1997) : « A Comparison of Alternative Sampling Methodologies for Census 2000 », *1997 Proceedings of the Section on Survey Research Methods, American Statistical Association*, à paraître.

MULRY, M.H. et SPENCER, B.D.(1991) : « Total Error in PES Estimates of Population(with discussion) », *Journal of the American Statistical Association, 86*, 839-863

STEFFEY, D.L. et BRADBURN, N.M.(Eds.)(1994) : *Counting People in the Information Age,* Panel to Evaluate Alternative Census Methods, Committee on National Statistics, National Research Council. Washington, D.C. : National Academy Press.

ESTIMATION DE VARIANCE EN PRÉSENCE D'IMPUTATION : OÙ EN SOMMES-NOUS ?

Eric Rancourt[1]

1. Introduction

Peu importe le temps et l'effort mis à la préparation d'une enquête, il y a toujours une partie des unités de l'échantillon pour laquelle l'information désirée n'est pas obtenue. Ce problème de non-réponse est souvent traité par l'imputation car elle permet d'utiliser à profit l'information partielle recueillie et de créer un ensemble de données complet. Plusieurs arguments peuvent être invoqués pour ou contre l'imputation mais le contexte du présent document est celui où l'imputation est considérée comme étant un fait accompli. Le lecteur trouvera une excellente discussion sur l'imputation dans Kovar et Whitridge (1995). Puisque l'on travaille à partir de données d'enquête, il s'ensuit que le processus d'estimation est entaché d'erreurs dont, entre autres, l'erreur échantillonnale. Afin de connaître la précision des estimations, on utilise habituellement le calcul de l'estimation de la variance. Si l'imputation est utilisée pour pallier la non-réponse, les estimations et leur précision seront affectées et ce, même avec la meilleure méthode d'imputation. On se doit donc de quantifier l'impact de l'imputation. Ce faisant, on pourra non seulement mieux informer les utilisateurs de la qualité réelle des estimations, mais on pourra également tenter de contrôler l'impact de l'imputation et de le réduire.

1. Eric Rancourt, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario) Canada, K1A 0T6.

Si on utilise les méthodes pour ensembles de données complets pour estimer la variance, on ne tient pas compte du fait que certaines données ont été imputées. C'est-à-dire que l'on fait l'hypothèse implicite que les données imputées se comportent comme des observations réelles. Dans ce cas, la variance totale sera sous-estimée. Plusieurs solutions à ce problème ont été développées et cet article décrit les suivantes :

1) l'imputation multiple, Rubin (1978, 1987)

2) l'approche assistée d'un modèle, Särndal (1990, 1992) et Deville et Särndal (1991) ;

3) l'approche à deux phases, Rao (1990) et Rao et Sitter (1995) ;

4) la technique du jackknife, Rao (1991) et Rao et Shao (1992) ;

6) la méthode pour imputation hot-deck, Provost (1995) ;

7) le bootstrap, Shao et Sitter (1996) ;

8) la méthode d'imputation de tous les cas, Montaquila et Jernigan (1997) ;

9) la méthode des échantillons balancés répétés (ou BRR), Shao, Chen et Chen (1998).

Cet article est divisé comme suit. À la section 2 on discute de l'importance de tenir compte de l'imputation dans le calcul de la variance. Par la suite, la section 3 contient une description des composantes de la variance totale. Dans la section 4 se trouvent les descriptions de ces diverses méthodes permettant le calcul correct de la variance, suivies d'une comparaison des méthodes à la section 5. On poursuit avec les points à considérer lors de l'application des méthodes à la section 6, et une présentation de l'approche adoptée à Statistique Canada à la section 7

2. Importance d'estimer correctement la variance

Lors de la production d'estimations à partir de données d'enquêtes, l'intérêt principal est évidemment pour les estimations ponctuelles. C'est pourquoi une importance mitigée est parfois accordée au calcul de la variance. Cependant, tel que mentionné dans Gagnon, Lee, Provost, Rancourt et Särndal (1997), l'estimation de la variance est très importante car elle permet :

1) de fournir une mesure de qualité des estimations ;

2) d'aider à tirer des conclusions correctes ;

3) aux agences statistiques d'informer les utilisateurs de la qualité des données.

Dans le cas de données imputées, il est encore plus important de mesurer la précision des estimations. En effet, l'imputation ajoute un processus supplémentaire pouvant être source d'erreurs. Ainsi, l'estimation de la variance qui tient compte de l'imputation permet, en plus des trois points cités plus haut :

4) de mieux connaître l'impact de l'imputation ;

5) d'améliorer l'estimation de la variance totale ;

6) d'effectuer une meilleure répartition des ressources entre un échantillon plus grand et un meilleur processus de vérification et imputation (selon la taille relative de la variance due à l'échantillonnage et la variance due à l'imputation).

Il est à noter que l'importance de tenir compte de l'imputation dans le calcul de la variance varie selon les conditions de l'enquête. Selon ces conditions (le taux de réponse, la méthode d'imputation, la fraction de sondage, la méthode d'estimation et la qualité des variables auxiliaires utilisées), la variance due à l'imputation pourrait atteindre un très grand pourcentage de la variance totale. Et même, dans le cas d'un recensement, il serait possible de calculer la variance due à l'imputation qui serait, par définition, 100% de la variance totale.

3. Imputation et estimation : Cadre théorique

À partir d'une population $U=\{1,\ldots,k,\ldots,N\}$, on tire un échantillon s. Le poids de sondage pour l'unité k est $a_k = 1/\pi_k$, où π_k est la probabilité de sélection. On désire estimer le total de la variable d'intérêt y, $Y_U = \sum_U y_k$. En l'absence de non-réponse, on utiliserait

$$\hat{Y}_s = \sum_s a_k g_k y_k$$

où, par exemple dans le cas de l'estimateur généralisé de régression (GREG) on a $g_k = 1 + (\sum_U \mathbf{x}_k - \sum_s \mathbf{x}_k / \pi_k)'(\sum_s \mathbf{x}_k \mathbf{x}_k' / \sigma_k^2 \pi_k)^{-1} \mathbf{x}_k / \sigma_k^2$ qui est le facteur d'ajustement au total de données auxiliaires, $\mathbf{X} = \sum_U \mathbf{x}_k$.

En présence de non-réponse, l'échantillon se retrouve scindé en deux parties : l'ensemble r des répondants et l'ensemble o des non-répondants. On a alors $o = s - r$. Pour traiter la non-réponse de l'unité k, $k \in o$, on impute une valeur

$\hat{y}_k$. Si cette imputation utilise une variable auxiliaire, elle sera dénotée par z_k. L'ensemble de données après imputation est $\{y_{\bullet k} : k \in s\}$, où

$$y_{\bullet k} = \begin{cases} y_k & \text{si } k \in r \\ \hat{y}_k & \text{si } k \in o. \end{cases}$$

Ainsi, en présence d'imputation, si la pondération demeure inchangée, l'estimation de Y_U devient

$$\hat{Y}_{\bullet s} = \sum_s a_k g_k y_{\bullet k}.$$

La quantité que l'on désire estimer étant Y_U, on peut décomposer l'erreur totale de la façon suivante :

$$\hat{Y}_{\bullet s} - Y_U = (\hat{Y}_s - Y_U) + (\hat{Y}_{\bullet s} - \hat{Y}_s),$$

où $\hat{Y}_s - Y_U$ est l'erreur d'échantillonnage et

$$\hat{Y}_{\bullet s} - \hat{Y}_s = \sum_o a_k g_k (\hat{y}_k - y_k)$$

est l'erreur d'imputation.

À partir des deux termes d'erreur ci-haut mentionnés, on pourra évaluer l'écart quadratique moyen. En supposant que l'imputation permette de faire une estimation sans biais, on aura plutôt l'expression de la variance. Voir par exemple Särndal (1992) pour la dérivation. L'expression de la variance totale que l'on obtient se compose de la variance due à l'échantillonnage, de la variance due à l'imputation et d'un terme mixte. Un estimateur de variance pour $\hat{Y}_{\bullet s}$ est donc

$$\hat{V}_{\text{TOT}} = \hat{V}_{\text{ÉCH}} + \hat{V}_{\text{IMP}} + \hat{V}_{\text{MIX}}$$

où $\hat{V}_{\text{ÉCH}}$ est l'estimateur de la variance échantillonnale, $\hat{V}_{\text{IMP}}$ est l'estimateur de la variance due à l'imputation et $\hat{V}_{\text{MIX}}$ correspond à deux fois la covariance entre les deux erreurs. Ce terme mixte est, dans plusieurs cas, relativement petit par rapport aux deux autres.

Dans le cas de l'échantillonnage aléatoire simple sans remise on pourrait, pour l'estimateur de la variance échantillonnale, $\hat{V}_{\text{ÉCH}}$, utiliser la formule habituelle d'estimation de variance,

$$\hat{V}_{\text{ORD}} = N^2 \frac{(1-f)}{n} \sum_s \frac{(y_{\bullet k} - \bar{y}_{\bullet s})^2}{n-1},$$

mais elle sous-estimerait la variance échantillonnale puisque les calculs se font sur l'ensemble de données après imputation. Il faudrait plutôt utiliser

$$\hat{V}_{\text{ÉCH}} = N^2 \frac{(1-f)}{n} \sum_s \frac{(y_k - \bar{y}_s)^2}{n-1},$$

mais y_k n'est pas disponible pour les non-répondants. Il faut donc estimer $\hat{V}_{\text{ÉCH}}$ par

$$\hat{V}_{\text{ÉCH}} = \hat{V}_{\text{ORD}} + \hat{V}_{\text{DIF}}$$

où $\hat{V}_{\text{DIF}}$ sera obtenu à l'aide d'un estimateur de $\hat{V}_{\text{ÉCH}} - \hat{V}_{\text{ORD}}$. Par exemple, dans Lee, Rancourt et Särndal (1995) on présente $\hat{V}_{\text{DIF}}$ pour l'imputation par la moyenne, par quotient, hot-deck et plus proche voisin. Pour certaines méthodes d'imputation, l'ensemble de données après imputation contient des valeurs ayant une variabilité suffisante pour que $\hat{V}_{\text{ORD}}$ soit assez près de $\hat{V}_{\text{ÉCH}}$. Il n'est donc pas nécessaire d'utiliser $\hat{V}_{\text{DIF}}$. Sous plusieurs conditions, c'est le cas de l'imputation par plus proche voisin et de l'imputation hot-deck. Dans Rao et Sitter (1997), on trouve une méthode d'imputation hot-deck construite spécifiquement pour éviter l'utilisation de $\hat{V}_{\text{DIF}}$.

On constate maintenant que, pour estimer correctement la variance totale, il ne suffit pas d'estimer la variance due à l'imputation, mais il faut également estimer correctement la variance due à l'échantillonnage.

4. Méthodes d'estimation de variance en présence d'imputation

Cette section décrit les méthodes d'estimation de variance en présence d'imputation mentionnées à la section 1. Pour simplifier la notation, le cas de l'échantillonnage aléatoire simple sans remise est traité.

4.1 Imputation multiple

La méthode de l'imputation multiple a été développée par Rubin (1978, 1987). Elle consiste à imputer, selon une méthode donnée, plusieurs valeurs pour chaque donnée manquante créant ainsi plusieurs ensembles de données après imputation $j = 1,...,J$. On peut donc utiliser les méthodes d'estimation habituelles sur chaque ensemble de données. Une fois que l'on a estimé la variabilité dans chaque ensemble de données, et entre les ensembles de données après imputation, on peut combiner les résultats et on a

$$\hat{V}_{\text{IM}} = \hat{V}_{\text{INTERNE}} + \hat{V}_{\text{ENTRE}}$$

qui ressemblent à $\hat{V}_{\text{ÉCH}}$ et $\hat{V}_{\text{IMP}}$. Plus spécifiquement, l'expression se lit comme suit:

$$\hat{V}_{\text{IM}} = \frac{1}{M}\sum_{j=1}^{M} N^2 \frac{1-f}{n} S^2_{y \bullet js} + \left(1+\frac{1}{M}\right)\frac{N^2}{M-1}\sum_{j=1}^{M}(\bar{y}_{\bullet js} - \bar{y}_{\bullet\bullet s})^2$$

où M est le nombre d'ensembles de données après imputation, $f = {}^{n}\!/_{N}$,

$$S^2_{y\bullet js} = \frac{1}{n-1}\sum\nolimits_s (y_{\bullet jk} - \bar{y}_{\bullet js})^2 \text{ et } \bar{y}_{\bullet\bullet s} = \frac{1}{M}\sum_{j=1}^{M}\bar{y}_{\bullet js}.$$

4.2 Approche assistée d'un modèle

L'approche assistée d'un modèle a été développée par Särndal (1990, 1992) et Deville et Särndal (1991, 1994). C'est une méthode pour l'imputation simple qui consiste à utiliser un modèle pour estimer la variance due à l'imputation en plus de celle due à l'échantillonnage. L'objectif est d'obtenir un estimateur de V_{TOT} pour

$\hat{Y}_{\bullet s}$ en construisant des estimateurs des composantes $V_{\text{ÉCH}}, V_{\text{IMP}}$, et V_{MIX} en utilisant un modèle de la forme

$$\xi : y_k = \beta z_k + \varepsilon_k \ ; \ E_\xi(\varepsilon_k) = 0; \ E_\xi(\varepsilon_k^2) = \sigma^2 z_k; \text{ et } E_\xi(\varepsilon_k \varepsilon_{k'}) = 0 \text{ pour } k \neq k',$$

où z est la variable d'imputation dont la valeur z_k est disponible au moins pour $k \in s$.

Les composantes $\hat{V}_{\text{ÉCH}}$, $\hat{V}_{\text{IMP}}$, et $\hat{V}_{\text{MIX}}$ doivent satisfaire $E_\xi(\hat{V}_{\text{ÉCH}} - V_{\text{ÉCH}}) = 0$, $E_\xi(\hat{V}_{\text{IMP}} - V_{\text{IMP}}) = 0$ et $E_\xi(\hat{V}_{\text{MIX}} - V_{\text{MIX}}) = 0$. En particulier, $\hat{V}_{\text{ÉCH}}$ est construit à l'aide de deux termes. Le premier consiste en la "formule ordinaire" de variance $\hat{V}_{\text{ORD}} = N^2 \frac{1-f}{n} S^2_{y\bullet s}$ calculée sur l'ensemble de données après imputation, avec $S^2_{y\bullet s} = \frac{1}{n-1} \sum_s \{y_{\bullet k} - (\sum_s y_{\bullet k} / n)\}^2$. On y ajoute le terme, $\hat{V}_{\text{DIF}}$, construit de façon à satisfaire $E_\xi\{\hat{V}_{\text{DIF}}\} = \frac{N^2(1-f)}{n} E_\xi\{S^2_{ys} - S^2_{y\bullet s}\}$. On obtient $\hat{V}_{\text{ÉCH}} = \hat{V}_{\text{ORD}} + \hat{V}_{\text{DIF}}$, et donc

$$\hat{V}_{\text{AM}} = \hat{V}_{\text{ORD}} + \hat{V}_{\text{DIF}} + \hat{V}_{\text{IMP}} + \hat{V}_{\text{MIX}}.$$

Par exemple, dans le cas de l'imputation par le quotient (ou ratio), où $\hat{y}_k = \hat{B} z_k$ avec $\hat{B} = \sum_r y_k / z_k$, on aura :

$$\hat{V}_{\text{ORD}} = N^2 \frac{1-f}{n} S^2_{y\bullet s} \qquad \hat{V}_{\text{DIF}} = N^2 \frac{1-f}{n^2} \sum_o z_k \hat{\sigma}^2$$

$$\hat{V}_{\text{IMP}} = \frac{N^2}{n^2} \sum_o z_k \left\{ \frac{\sum_o z_k}{\sum_r z_k} + 1 \right\} \hat{\sigma}^2 \qquad \hat{V}_{\text{MIX}} = N^2 \frac{(1-f)}{n^2} \sum_o z_k \left\{ \frac{\sum_o z_k}{\sum_r z_k} - 1 \right\} \hat{\sigma}^2,$$

avec $\hat{\sigma}^2 = \sum_r e_k^2 / \sum_r z_k$ et $e_k = y_k - \hat{B} z_k$.

4.3 Approche en deux phases

Développée par Rao (1990) et Rao et Sitter (1995) à partir de l'idée de l'échantillonnage à deux phases, cette méthode suppose que l'échantillon est la première phase d'un plan de sondage, et que les répondants constituent l'échantillon de deuxième phase. De façon implicite, on suppose donc que les répondants forment un échantillon aléatoire des unités de l'échantillon. L'idée de base est d'obtenir un estimateur de variance formé d'un terme de variance due à la première phase et d'un terme issu de la deuxième phase. On a donc, dans le cas de l'imputation par le ratio

$$\hat{V}_{\text{DP1}} = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_{yr}^2 + N^2\left(\frac{1}{m}-\frac{1}{N}\right)S_{er}^2.$$

où $S_{yr}^2 = \sum_r (y_k - \bar{y}_r)^2/(m-1).$

Pour utiliser l'information auxiliaire, Rao (1990) et Rao et Sitter (1995) suggèrent plutôt l'expression suivante basée sur la même approche :

$$\hat{V}_{\text{DP}} = N^2\left(\frac{1}{n}-\frac{1}{N}\right)\hat{B}^2 S_{zs}^2 + 2N^2\left(\frac{1}{n}-\frac{1}{N}\right)\hat{B}S_{zer} + N^2\left(\frac{1}{m}-\frac{1}{N}\right)S_{er}^2.$$

où $S_{zer} = \sum_r e_k z_k /(m-1).$

4.4 Technique du Jackknife

La technique du jackknife a pour principe de recalculer l'estimateur après avoir enlevé une ou plusieurs unités de l'échantillon. La variance entre les estimations obtenues est utilisée pour obtenir une estimation de la variance de l'estimateur calculé sur l'ensemble de l'échantillon. Lorsque l'unité j est enlevée, l'estimateur du total Y_U est donné par $\hat{Y}_{\bullet s}^{(j)} = N\sum_{k \neq j \in s} y_{\bullet k} /(n-1)$. L'estimateur par le jackknife est $\hat{V} = \frac{n-1}{n}\sum_{j \in s}(\hat{Y}_{\bullet s}^{(j)} - \hat{Y}_{\bullet s})^2$.

En présence d'imputation, le jackknife, tel que définit ci-haut, sous-estime V_{TOT}. Pour cette situation, Rao and Shao (1992) proposent un estimateur de variance qui

corrige le jackknife en ajustant les valeurs imputées lorsque l'unité enlevée fait partie de l'ensemble des répondants. L'ensemble de données ajustées est donné par

$$y_{\bullet k}^{(aj)} = \begin{cases} y_k & \text{si } k \in r \\ \hat{y}_k + a_k^{(j)} & \text{si } k \in o \text{ et } j \in r \\ \hat{y}_k & \text{si } k \in o \text{ et } j \in o \end{cases}$$

où $y_{\bullet k}^{(aj)}$ est la valeur imputée ajustée et $a_k^{(j)}$ est l'ajustement. L'estimateur de variance par le jackknife est alors

$$\hat{V}_{\text{JK}} = \frac{n-1}{n} \sum_{j \in s} (\hat{Y}_{\bullet s}^{(aj)} - \hat{Y}_{\bullet s}^{(a)})^2$$

où $\hat{Y}_{\bullet s}^{(aj)} = \frac{N}{n-1} \sum_{k \neq j \in s} y_{\bullet k}^{(aj)}$ et $\hat{Y}_{\bullet s}^{(a)} = \frac{1}{n} \sum_{j \in s} \hat{Y}_{\bullet s}^{(aj)}$. Cet estimateur fonctionne bien lorsque la correction pour population finie (cpf), $1-f$ avec $f = n/N$, n'est pas nécessaire. Pour les situations où la cpf est requise, Lee, Rancourt et Särndal (juillet 1995) proposent la correction suivante à l'estimateur de variance par le jackknife $\hat{V}_{\text{JK}}^* = \hat{V}_{\text{JK}} - N\hat{S}_{yU}^2$, où $\hat{S}_{yU}^2$ est un estimateur sans biais de S_{yU}^2.

Les ajustements $a_k^{(j)}$ dépendent de la méthode d'imputation. On peut trouver les ajustements pour l'imputation par la moyenne et hot-deck dans Rao (1991), et Rao et Shao (1992), pour l'imputation par quotient dans Rao (1991) et Rao et Sitter (1995), et pour l'imputation par plus proche voisin dans Kovar and Chen (1994). Il est également intéressant de noter que Rao et Sitter (1995) présentent une linéarisation de l'estimateur de variance par la méthode du jackknife.

4.5 Bootstrap

La technique du bootstrap pour l'estimation de la variance en présence de données imputées a été développée par Shao et Sitter (1996). De même que la technique habituelle du bootstrap, elle consiste à tirer plusieurs échantillons à partir de l'échantillon d'origine en reproduisant la méthode d'échantillonnage. Cependant, dans le cas de données imputées, chaque donnée imputée se retrouvant dans l'échantillon bootstrap doit être ré-imputée par la même procédure d'imputation ayant servi à l'origine. Ainsi, l'estimateur de la variance par bootstrap est

$$\hat{V}_{\text{BOOT}} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{Y}_{*k}^{(b)} - \overline{Y}_{*s}\right)^2$$

où B est le nombre d'échantillons bootstrap, $\hat{Y}_{*k}^{(b)}$ le total pour l'échantillon b estimé sur les données après ré-imputation, et $\overline{Y}_{*k}$ est la moyenne des totaux sur tous les échantillons bootstrap. Cette formule présuppose un échantillonnage avec remise. Dans le cas de l'échantillonnage sans remise, Shao et Sitter (1996) décrivent trois méthodes qui peuvent être employées.

4.6 Méthode pour Hot-deck

Dans le cas de l'imputation hot-deck, il existe une approche basée sur le plan de sondage qui permette d'estimer correctement la variance totale pour un mécanisme de réponse uniforme. Cette approche à été développée par Provost (1995). Elle est basée sur le fait que l'imputation hot-deck correspond à un tirage aléatoire simple d'une unité parmi les m répondants, conditionnellement à l'échantillon s et à l'ensemble de répondants r. Pour estimer la variance totale, on doit donc estimer la variance due à l'échantillonnage et la variance due à l'imputation. On a alors

$$\hat{V}_{\text{HD}} = \hat{V}_{\text{ÉCH}} + \hat{V}_{\text{IMP}}.$$

Pour l'échantillonnage aléatoire simple sans remise et un mécanisme de réponse uniforme, les deux termes sont

$$\hat{V}_{\text{ÉCH}} = \frac{mn(n-1)}{(n^2 - n + m)(m-1)} N^2 \frac{(1-f)}{n} \sum_s \frac{(y_{\bullet k} - \overline{y}_{\bullet s})^2}{n-1}$$

$$\hat{V}_{\text{IMP}} = \frac{(n^2 + m - n - m^2)(n-1)}{(n^2 - n + m)(m-1)} N^2 \frac{1}{n} \sum_s \frac{(y_{\bullet k} - \overline{y}_{\bullet s})^2}{n-1}.$$

4.7 Méthode d'imputation de tous les cas

La technique d'imputation de tous les cas consiste à imputer une valeur à toutes les unités de l'échantillon y compris les répondants. L'estimation ponctuelle s'effectue alors en utilisant les données imputées de tout l'échantillon. Pour le calcul de la variance, on dispose donc de résidus qui vont permettre d'évaluer l'erreur d'imputation, et donc d'estimer la variance due à l'imputation. Cette méthode,

décrite dans Montaquila et Jernigan (1997) a été développée récemment pour l'imputation par donneur dans le cas de populations infinies.

4.8 La Méthode des échantillons balancés répétés (ou BRR)

L'estimation de variance en présence d'imputation a aussi été développée pour la méthode des échantillons balancés répétés. Elle consiste en un ajustement des données imputés similaire à celui pour le jackknife, pour chacun des échantillons balancés. Une fois ces ajustements effectués, on peut utiliser la formule

$$\hat{V}_{\text{BRR}} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}_{\bullet k}^{(r)} - \overline{Y}_{\bullet}\right)^2,$$

où R est le nombre d'échantillons balancés, et $\hat{Y}_{\bullet}^{(r)}$ est l'estimateur pour l'échantillon r, calculé avec les ajustements.

La méthode est décrite dans Shao, Chen et Chen (1998), qui paraîtra bientôt.

5. Comparaisons des méthodes

Plusieurs caractéristiques différencient les méthodes d'estimation de variance présentées à la section précédente. La discussion qui suit aborde sept thèmes qui sont ensuite résumés dans un tableau général. Les termes entre parenthèses sont utilisés dans le tableau 1 à la section 5.8 pour référer aux différentes caractéristiques des méthodes.

5.1 Possibilité d'obtenir une estimation de $V_{ÉCH}$ et de V_{IMP} (V_{IMP})

Dans toutes les méthodes sauf le jackknife, le bootstrap et le BRR, on obtient différents termes qui représentent plus ou moins bien $\hat{V}_{\text{ÉCH}}$ et $\hat{V}_{\text{IMP}}$. Il peut être avantageux de disposer de cette décomposition de l'estimation de la variance totale pour mieux connaître l'impact de l'imputation. En connaissant l'importance relative de $\hat{V}_{\text{ÉCH}}$ et $\hat{V}_{\text{IMP}}$ par rapport à $\hat{V}_{\text{TOT}}$, on dispose ainsi d'une mesure qui pourrait permettre de mieux répartir, dans les enquêtes répétées, les ressources entre deux

objectifs importants, soient l'amélioration du plan de sondage ou du système d'imputation.

5.2 Nombre d'imputations requises pour chaque unité (# Imp)

Une caractéristique qui est propre aux méthodes d'estimation de variance en présence d'imputation est le nombre d'ensembles de données créés. C'est-à-dire que pour l'imputation simple, il n'y a qu'un seul ensemble de données alors qui pour l'imputation multiple, il y en a plusieurs. Ce surplus d'ensembles à entreposer et à maintenir peut engendrer des coûts supplémentaires. Par contre, l'imputation multiple permet l'utilisation des méthodes habituelles d'estimation de variance. Il est ainsi possible de les appliquer sur chaque ensemble de données pour ensuite combiner les résultats afin d'obtenir une estimation de la variance totale qui inclut la variance due à l'imputation. Il est aussi à noter que le bootstrap est en fait une technique d'imputation multiple, puisque le processus d'imputation doit être répété à chaque itération du bootstrap.

5.3 Nécessité d'identifier les répondants et les non-répondants (Flag)

Pour les méthodes d'estimation de variance qui s'appliquent à l'imputation simple, il est nécessaire de savoir si les données de l'échantillon font partie de l'ensemble des répondants ou de l'ensemble des non-répondants. En d'autres termes, on doit disposer d'identificateurs d'imputation pour chaque unité de l'échantillon. De plus, on doit également connaître la méthode d'imputation utilisée pour obtenir la bonne formule d'estimation de variance et pour utiliser la bonne correction pour la technique du jackknife.

5.4 Restriction sur les méthodes d'imputation (Méthode)

Toutes les méthodes d'estimation de variance ont évidemment des limites, mais certaines ont été développées pour des méthodes d'imputation spécifiques. C'est le cas de la méthode hot-deck qui, comme son nom l'indique, est conçue pour l'imputation hot-deck et de la méthode d'imputation de tous les cas qui s'applique présentement au cas de l'imputation par donneur. En ce qui concerne les autres méthodes, elles semblent pouvoir s'appliquer à de plus grandes familles de méthodes d'imputation. Également, la méthode de l'imputation multiple requiert que la méthode d'imputation soit « propre » au sens décrit dans Rubin (1987).

5.5 Hypothèses sur le mécanisme de non-réponse (Non-rép)

Les mécanismes de non-réponse peuvent être classés en trois groupes, comme décrit dans Rubin (1976) et dans Rancourt, Lee et Särndal (1994) :

1) Mécanismes où les données sont manquantes de façon aléatoire (MCAR[2]. ou uniforme) où la non-réponse ne dépend d'aucune variable de l'échantillon et est distribuée uniformément.

2) Mécanismes où les données sont manquantes de façon aléatoire (MAR ou non-confondu) conditionnellement à une ou plusieurs variables auxiliaires. Dans ce cas, la non-réponse peut dépendre d'une variable auxiliaire mais ne dépend pas de la variable d'intérêt.

3) Mécanismes où les données sont manquantes de façon non aléatoire (NMAR ou confondu) où la non-réponse dépend de la variable d'intérêt.

Toutes les méthodes décrites dans cet article fonctionnent dans le cas 1 et sont vulnérables au mécanisme de réponse dans le cas 3. Dans le cas 2, les méthodes utilisant un modèle (et donc des variables auxiliaires) et les méthodes de ré-échantillonnage se comportent assez bien selon les conditions de l'enquête. Par contre, une méthode comme l'approche en deux phases suppose que l'ensemble des répondants soit un échantillon aléatoire simple de l'échantillon. La méthode est donc très bien adaptée à la situation 1 mais n'est pas tellement robuste aux situations 2 et 3.

5.6 Utilisation d'un modèle (Modèle)

Les méthodes d'imputation peuvent toutes être exprimées de façon explicite ou implicite par un modèle. Si le modèle supposé pour l'imputation n'est pas bon, l'estimation de variance sera moins exacte. Par contre, si le modèle est bon, l'estimation de la variance sera sans biais. Un modèle peut être utilisé de façon implicite ou explicite dans le développement d'une approche d'estimation de variance. Seule la méthode assistée d'un modèle en fait une utilisation explicite.

---

2. Les acronymes MCAR, MAR et NMAR proviennent des termes anglais : « Missing Completely At Random », « Missing At Random », et « Not Missing At Random ».

5.7 Nature des utilisateurs des données après imputation (À qui)

Après la diffusion des résultats d'une enquête, ce sont les clients se procurant les données qui seront les utilisateurs. Par contre, en ce qui a trait au calcul de la variance, on peut distinguer deux types d'analystes de données :

1)Les « imputeurs », qui font eux-mêmes l'analyse et fournissent ensuite les estimations et une mesure de leur précision aux clients.

2)Les clients, qui font le calcul des mesures de précision lors de leurs analyses, après avoir reçu des « imputeurs » l'ensemble de données après imputation.

Dans le premier cas, il n'est pas nécessaire de construire plusieurs ensembles de données pouvant être analysées à l'aide de logiciels simples, puisque l'imputeur dispose des connaissances et de toute l'information requise pour évaluer la précision des estimations. Dans le deuxième cas, il est par contre important de fournir les outils dont l'analyste a besoin pour effectuer ses calculs. C'est dans ce dernier cas, que l'imputation multiple apparaît comme une excellente solution. En effet, une fois les multiples ensembles de données créés, l'analyste n'a qu'à répéter son travail sur chacun des ensembles de données, pour ensuite combiner les résultats selon la formule donnée à la section 4.1. Par contre, l'analyste doit être disposé à prendre le temps de répéter son analyse et le stockage de J ensembles de données n'est pas intéressant.

5.8 Différences entre les méthodes

Le tableau 1 présente un résumé de chacune des caractéristiques mentionnées ci-haut. Pour chaque caractéristique, les entrées sont les suivantes :

Vimp :	Peut-on séparer $\hat{V}_{ÉCH}$ et $\hat{V}_{IMP}$?	(Oui, Non)
#Imp :	Nombre d'imputations requises :	(1, >1)
Flag :	Besoin d'identifier répondants et non-répondants?	(Oui, Non)
Méthodes :	Restriction sur la méthode d'imputation :	(Méthode)
Non-rép :	Hypothèse sur la non-réponse :	(Uniforme, confondu)
Modèle :	La méthode utilise-t-elle explicitement un modèle?	(Oui, Non)
À qui :	Utilisateurs de l'ensemble de données :	(Interne, Externe)

Tableau 1
Caractéristiques des méthodes d'estimation de variance

	Vimp	#imp	Flag	Méthodes	Non-rép.	Modèle	À qui
Imputation Multiple	Oui	>1	Non	Propre	Pas confondu	Non	Externe
Assistée Modèle	Oui	1	Oui		Pas confondu	Oui	Interne
Jackknife	Non	1	Oui		Pas confondu	Non	Interne
2 phases	Oui	1	Oui		Uniforme	Non	Interne
Hot-deck	Oui	1	Oui	Hot-deck	Pas confondu	Non	Interne
Bootstrap	Non	>1	Oui		Pas confondu	Non	Interne
Tous les cas	Oui	1	Oui	Donneur	Pas confondu	Non	Interne
BRR	Non	1 ou >1	Oui		Pas confondu	Non	Interne

Pour plus de renseignement sur les différences, voir Lee, Rancourt et Särndal (1994) et Kovar et Chen (1994) qui ont effectué des expériences de Monte-Carlo et comparé certaines de ces méthodes sous différents modèles de population et types de non-réponse.

5.9 Disponibilité des méthodes

Il n'y a pas encore de logiciel d'estimation qui offre vraiment la possibilité de tenir compte de l'imputation dans l'estimation de la variance. Cependant, pour un système comme le Système généralisé d'estimation (SGE) de Statistique Canada décrit dans Estevao, Hidiroglou et Särndal (1995), on présente dans Lee, Rancourt et Särndal (août 1995) les éléments de base de l'estimation en présence d'imputation. De plus, des développements sont en cours à Statistique Canada sur un prototype appelé SIMPVAR afin de tenir compte de l'imputation. Ce système est brièvement décrit à la section 7.

6. Points à considérer lors de la mise en oeuvre

L'estimation de variance, et plus particulièrement dans le cas où il y a eu de l'imputation, est un problème souvent considéré seulement vers la fin du traitement des données. On devrait plutôt considérer ce problème dans son ensemble. C'est-à-dire que, dès le développement de l'enquête, il faut planifier l'estimation de

variance de façon à ne pas se retrouver face à un manque d'information pouvant être nécessaire au calcul d'une estimation de variance qui tient compte de l'imputation. La liste (non exhaustive) suivante contient des points qui peuvent contribuer à simplifier ou même à rendre possible l'estimation de variance. Plus de détails sont fournis dans Rancourt (1996).

1) L'imputation et l'estimation ne doivent pas être considérées comme deux approches séparées, mais comme deux sous-étapes d'un seul et même processus.

2) La création des groupes d'imputation devrait tenir compte (et se rapprocher) autant que possible des domaines d'estimation.

3) Les méthodes d'imputation utilisant de l'information auxiliaire contribuent à une robustesse accrue face aux différents mécanismes de non-réponse possibles.

4) Les groupes d'imputation ne doivent pas être basés uniquement sur les combinaisons possibles des variables catégoriques, mais également sur l'adéquation du modèle d'imputation.

5) L'utilisation des méthodes d'imputation ayant une composante stochastique permet de préserver les distributions et facilite le calcul de la variance. Voir par exemple Rao et Sitter (1997).

6) Des identificateurs (flags) de réponse et de méthode d'imputation doivent être assignés et conservés.

7) Il est préférable de restreindre le nombre de méthodes d'imputation utilisées à l'intérieur du même groupe d'imputation pour préserver la cohérence des données. Cette situation est abordée dans Rancourt, Lee et Särndal (1993).

8) Non seulement les méthodologistes / statisticiens doivent participer à l'élaboration de l'imputation et de l'estimation, mais il est essentiel d'y faire participer les spécialistes du sujet de l'enquête.

9) La stratégie globale doit être simple.

7. Estimation de variance en présence d'imputation à Statistique Canada

Cette section présente un aperçu des caractéristiques du Système généralisé d'estimation (SGE) de Statistique Canada. Ce système est conçu pour le cas d'ensembles de données complets, mais les plans de développement prévoient l'incorporation de méthodes d'estimation de variance tenant compte de l'imputation.

À cette fin, un prototype, SIMPVAR (Système pour tenir compte de l'imputation dans l'estimation de la variance), est en développement et est décrit ci-après.

7.1 Le Système généralisé d'estimation (SGE) de Statistique Canada

Le Système généralisé d'estimation (SGE), Estevao, Hidiroglou et Särndal (1995) est une application du progiciel SAS développée à Statistique Canada afin de produire des estimations par domaines pour un large éventail de situations. Plusieurs enquêtes font utilisation du SGE pour les calculs de totaux, de moyennes ou de ratios. Le système a été construit afin de satisfaire plusieurs plans de sondage et il fournit un vaste choix d'estimateurs à travers l'utilisation de l'estimateur de régression généralisé ou GREG, Särndal, Swensson et Wretman (1992). De plus, le calcul de la variance peut s'effectuer à l'aide de la formule issue de la linéarisation de Taylor ou de la technique du jackknife.

7.2 Le Système pour tenir compte de l'imputation dans l'estimation de variance (SIMPVAR)

Le SGE a été développé pour le cas d'ensembles de données complets, c'est-à-dire sans imputation. Pour le cas de données après imputation, l'approche assistée d'un modèle présentée à la section 4.2 est présentement en développement et a été incorporée dans un système appelé SIMPVAR. Les bases de ce système ont été jetées dans Gagnon, Lee, Rancourt et Särndal (1996). Éventuellement, ce système sera intégré au SGE. La version courante de SIMPVAR est très conviviale et fonctionne avec un système de menus. La plupart des conditions présentes dans le SGE peuvent être traitées dans SIMPVAR ; il suffit de fournir un identificateur pour les répondants et les non-répondants, d'indiquer la méthode d'imputation, d'identifier s'il y a lieu la variable auxiliaire utilisée pour l'imputation et d'indiquer le donneur pour l'imputation par donneur.

L'essence de SIMPVAR est de calculer la variance due à l'imputation et de l'ajouter à la variance due à l'échantillonnage, qui aurait été calculée au préalable par un autre système. On a donc la formule

$$\hat{V}_{\text{TOT}} = \hat{V}_{\text{ÉCH}} + \hat{V}_{\text{IMP}},$$

où $\hat{V}_{\text{IMP}}$ est obtenu par SIMPVAR et $\hat{V}_{\text{ÉCH}}$ par le SGE ou un autre système.

Ainsi SIMPVAR peut être perçu comme un module qui ajoute simplement une colonne (variance due à l'imputation) au fichier final contenant les estimations et qui en modifie une autre (variance totale).

8. Conclusion

Le problème d'estimation de la variance en présence d'imputation en est un qui mérite l'attention. Plusieurs méthodes ont été développées et peuvent être utilisées dans les enquêtes. À Statistique Canada, la méthode assistée d'un modèle a été retenue et est présentement en développement. Elle existe dans un système développé sous la forme d'un prototype appelé SIMPVAR et sera éventuellement incorporée dans le Système généralisé d'estimation (SGE). Plusieurs enquêtes pourront alors l'utiliser pour estimer la variance due à l'imputation, et donc mieux connaître la précision des estimations. Ceci permettra de rendre plus efficace l'assignation des ressources pour l'imputation en plus de pouvoir informer avec plus de précision les utilisateurs sur la qualité des données.

DEVILLE, J.-C., SÄRNDAL, C.-E., « Estimation de la variance en présence de données imputées », *Proceedings of Invited Papers for the 48th Session of the International Statistical Institute*, Book 2, Subject 17, 3e17, 1991.

DEVILLE, J.-C., SÄRNDAL, C.-E., « Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator », *Journal of Official Statistics*, 10, 381-394, 1994.

ESTEVAO, V., HIDIROGLOU, M.A., SÄRNDAL, C.-E., « Methodological principles for a generalized estimation system at Statistics Canada », *Journal of Official Statistics*, 11, 181-204, 1995.

GAGNON, F., LEE, H., RANCOURT, E., SÄRNDAL, C.-E., « Estimating the variance of the Generalized regression estimator in the presence of imputation for the Generalized Estimation System », *Recueil de la section de méthodologie d'enquête*, Société Statistique du Canada, 151-156, juin 1996.

GAGNON, F., LEE, H., PROVOST, M., RANCOURT, E., SÄRNDAL, C.-E., « Estimation de la variance en présence d'imputation », *Recueil du Symposium 97 : Nouvelles orientations pour les enquêtes et les recensements,* à paraître, Statistique Canada, Ottawa, novembre 1997.

KOVAR, J.G., CHEN, E., « Méthode du jackknife pour l'estimation de la variance en présence de données imputées », *Technique d'enquête*, 20, 47-55, 1994.

KOVAR, J.G., WHITRIDGE, P.J., « Imputation of Business Survey Data », *Business Survey Methods,* Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. et Kott, P.S. (éditeurs), 403-423, New York: John Wiley and Sons, 1995.

LEE, H., RANCOURT, E., SÄRNDAL, C.-E., « Experiment with Variance Estimation from Survey Data with Imputed Values », *Journal of Official Statistics*, 10, 231-243, 1994.

LEE, H., RANCOURT, E., SÄRNDAL, C.-E., « Jackknife Variance Estimation for Data with Imputed Values », *Recueil de la section de méthodologie d'enquête,* 111-115, Société Statistique du Canada, juillet 1995.

LEE, H., RANCOURT, E., SÄRNDAL, C.-E., « Variance estimation in the presence of imputed data for the Generalized Estimation System », *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 384-389, août 1995.

MONTAQUILA, J. M., JERNIGAN, R. W., «Variance Estimation in the Presence of Imputed data», *Proceedings of the Section on Survey Research Methods,* American Statistical Association, à paraître, août 1997.

PROVOST, M., *Estimation de la variance dans les sondages utilisant l'imputation hot-deck.* Mémoire de maîtrise, Université de Montréal, 1995.

RANCOURT, E., « Issues in the Combined Use of Statistics Canada's Generalized Edit and Imputation System and Generalized Estimation System », *Survey and Statistical Computing : Proceedings of The Second ASC International Conference,* Association for Survey Computing, 185-194, septembre 1996.

RANCOURT, E., LEE, H., SÄRNDAL, C.-E., « Variance Estimation under More Than one Imputation Method », *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, 374-379, juin 1993.

RANCOURT, E., LEE, H., SÄRNDAL, C.-E., « Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Nonresponse », *Survey Methodology*, 20, 137-147, 1994.

RAO, J.N.K., « Variance Estimation under Imputation for Missing Data ». Rapport technique, Statistique Canada, Ottawa, 1990.

RAO, J.N.K., « Jackknife Variance Estimation under Imputation for Missing Data. Rapport technique, Statistique Canada, Ottawa, 1991.

RAO, J.N.K., SHAO, J., « Jackknife variance estimation with survey data under hot-deck imputation ». *Biometrika*, 79, 811-822, 1992.

RAO, J.N.K., SITTER, R.R., « Variance estimation under two-phase sampling with application to imputation for missing data », *Biometrika*, 82, 453-460, 1995.

RAO, J.N.K., SITTER, R.R., « Efficient Random Imputation for Missing Data in Complex Surveys », Rapport technique, Carleton University et Simon Fraser University, 1997.

RUBIN, D.B., « Inference and missing data », *Biometrika*, 63, 581-590, 1976.

RUBIN, D.B., « Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse ». *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 20-34, 1978.

RUBIN, D.B., *Multiple imputation for nonresponse in surveys*, New York: John Wiley and Sons, 1987.

SÄRNDAL, C.-E., « Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation », *Recueil du Symposium '90: Mesure et amélioration de la qualité des données*, 337-347. Statistique Canada, Ottawa, 1990.

SÄRNDAL, C.-E., « Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation », *Techniques d'enquête*, 18, 257-268, 1992.

SÄRNDAL, C.-E., « For a Better Understanding of Imputation », *Proceedings of the 6th Workshop on Household Survey Nonresponse,* Helsinki, Octobre 1995.

SÄRNDAL, C.-E., SWENSSON, B., WRETMAN, J.H., *Model Assisted Survey Sampling*. New York: Springer-Verlag, 1992.

SHAO, J., CHEN, Y, CHEN, Y., « Balanced Repeated Replication for Stratified Multistage Survey Data under Imputation », *Journal of the American Statistical Association*, à paraître, 1998.

SHAO, J., SITTER, R.R., « Bootstrap for imputed survey data », *Journal of the American Statistical Association*, 91, 1278-1288, 1996.

Session 1

Le panel européen de ménages

LES ENQUETES PAR PANEL : EN QUOI DIFFERENT-ELLES DES AUTRES ENQUETES ?
suivi de : comment attraper une population en se servant d'une autre

Jean-Claude Deville

Le terme "*panel*" est souvent utilisé de façon abusive. Dans certains milieux un peu snob de la pub, il est utilisé pour échantillon. En économétrie, dans la locution "données de panel", il se réfère à des modèles à deux indices dont l'un est ordonné. On appréciera au passage l'usage délicieux du mot "données" : données par qui ? comment ? Le but de cet exposé est de montrer les contraintes logiques liées à l'élaboration de ces fameuses "données".

Un minimum de clarifications s'impose. Après quelques constatations de bon sens sur la notion de temps et d'identification, on essaiera de passer en revue les différentes formes de collecte où le temps intervient : enquêtes répétées, enquêtes de cohortes, échantillons rotatifs ou coordonnés, enquêtes continues, panels.

On introduira un formalisme unificateur de toutes ces formes d'enquête. Vu sous un certain angle, les panels apparaissent comme des enquêtes ordinaires justiciables des méthodes les plus habituelles d'analyse, d'estimation, d'estimation de variance et même d'échantillonnage.

Les exemples seront pris de préférence dans le domaine des enquêtes auprès des personnes, mais sont transposables aux enquêtes auprès des autres agents économiques. On verra en particulier pourquoi la notion de panel de ménage est à peu près vide de sens, bien qu'on puisse assez facilement faire des statistiques de ménages à l'aide d'un échantillon de personnes.

1 - Population et identifiant

On appelle généralement panel (cf. [LAVALLEE (1996)]) toute enquête où les unités sont enquêtées à plusieurs dates successives. Cette définition, qui ne prend en compte que le processus de collecte des données, reste tout à fait imprécise si nous voulons examiner les problèmes statistiques liés à l'introduction du temps, c'est-à-dire de variables datées, dans les enquêtes par sondage.

Il faut donc retourner à la racine des choses. La période d'étude T est une intervalle de temps $T = (t_0, t_f)$ où t_f est généralement fini (c'est l'horizon de l'étude) mais souvent, en pratique, indéfini et donc pris égal à l'infini (ce qui ne manque pas de poser des problèmes angoissants, comme chaque fois qu'on est, dans la pratique, confronté à ce mystère). A chaque instant t est associée une population finie identifiée U_t, c'est-à-dire l'objet potentiel d'un sondage. A chaque individu k (k est l'"étiquette", le "label" ou l'identifiant) de U_t est associé un vecteur de variables d'intérêt $y_k^{(t)}$, un vecteur de variables auxiliaires $x_k^{(t)}$; éventuellement on dispose aussi d'information auxiliaire externe $\mathcal{Z}^{(t)}$.

Un système d'enquêtes répétées consiste à réaliser à diverses dates $t \in F = \{t_1, t_2, \ldots, t_n, \ldots\}$ (où F est fini (dans tout intervalle fini de T dans le cas où $t_f = \infty$!)), des enquêtes par sondage indépendantes les unes des autres dans les $U_t (t \in F)$ destinées à recueillir les "mêmes" variables. C'est la pratique traditionnelle des recensements successifs de la population (au ¼ !) ou des enquêtes-logements. Le but de l'exploitation statistique est alors de fixer des niveaux de y aux différentes périodes et de les comparer pour évaluer des évolutions globales. Dans le cas où certaines unités statistiques auxiliaires sont pérennes (zone d'emploi, communes), le cumul de plusieurs enquêtes successives peut aider à élaborer des statistiques "locales". Il n'y a pas grand chose à dire de plus sur cette façon de traiter le temps, sinon qu'il ne joue aucun rôle particulier : on pourrait dire que t désigne un pays européen par exemple. Autrement dit on ne suppose absolument rien - dans la technique d'enquête en tout cas - sur le lien entre les populations $U_t (t \in F)$.

La spécificité du temps (implicite mais qu'il vaut mieux expliciter) est que la population U_t varie en un certain sens "continûment". Soyons un peu formel : soit $U = \bigcup_{t \in T} U_t$ la population d'étude. On dira que la population est "renouvelée" si U est fini (ou, si T est non borné, si $U_B = \bigcup_{t \in B} U_t$ est fini pour toute partie bornée B de T, (on a déjà des problèmes avec l'infini !)), autrement dit si une étiquette k de U figure dans une partie T_k assez grosse. On ne restreindra guère la généralité en supposant que T_k est un intervalle (ou, éventuellement, une réunion finie d'intervalles *(Exemple : Etudes sur les chômeurs)*).

Ceci suppose implicitement qu'il existe un moyen d'identifier les individus de U, les unités statistiques, comme étant les mêmes à deux époques distinctes et donc au cours de toute la période d'étude. Cette condition, théorique et pratique, est essentielle, fondatrice, dès qu'on veut aller au delà des études par enquêtes répétées et qu'on veut travailler sur des données dites longitudinales.

Cette remarque a des conséquences immédiates sur la façon d'envisager la statistique longitudinale des populations humaines. Il paraît à peu près clair qu'on puisse identifier d'une façon rigoureuse les personnes. Sans aller jusqu'à des arguments génétiques, l'usage du NIR ou simplement du "nom, prénom, date et lieu de naissance" semble faire l'affaire. (Ceci dit, on peut se poser des questions sur des époques ou des pays où on ignore les tests génétiques et où les inscriptions d'Etat-Civil sont un tantinet déficientes !). Encore faut-il pouvoir, techniquement, utiliser cet identifiant. Ce n'est par exemple guère possible de façon massive dans les recensements ; ceux-ci autorisent néanmoins la confection de l'Echantillon Démographique Permanent (E.D.P, "Le Panel Démographique" pour les initiés").

L'identification des <u>ménages</u> pose un tout autre problème qui n'a vraisemblablement pas de solution exempte d'arbitraire. Stricto sensu, un ménage est un ensemble de personnes et s'identifie par la liste de ses membres. Un ménage est donc le même à t et t' s'il est composé des mêmes personnes.

Cette définition a l'avantage de prendre en compte les déménagements. En revanche toute arrivée ou tout départ signifie la "mort" du ménage et la "naissance" d'un autre, même en cas de "mouvement naturel" au sens démographique habituel : "naissances" ou "décès" c'est-à-dire entrées et sorties de la population d'étude (ces notions peuvent se formaliser facilement de façon rigoureuse, mais nous n'insisterons pas ici là-dessus).

Le "solde naturel" de la population entre t et $t'\left(t < t'\right)$ est la différence symétrique $S_{t,t'} = U_t \Delta U_{t'} = \left\{k ; \left(k \in U_t \text{ et } k \notin U_{t'}\right) \text{ (décès) ou } \left(k \notin U_t \text{ et } k \in U_{t'}\right) (\text{naissances})\right\}$. On peut dire qu'un ménage reste identique à lui-même, et donc peut conserver le même identifiant, s'il n'est affecté que par le mouvement naturel. Cette notion n'est pas si simple à formaliser, et donc à vérifier concrètement dans une opération d'enquête : elle est en effet différentielle en ce sens qu'elle demande la prise en compte de tous les "<u>événements démographiques</u>" survenus ente t et t' et mettant en cause les individus susceptibles d'avoir appartenu à ce ménage longitudinal au cours de la période. Identifier les ménages m_t et $m_{t'}$ lors de deux recensements, par exemple, est impossible même si la population $S_{t,t'}$ est parfaitement connue ! La seule définition possible devrait contenir la clause : $m_t \Delta m_{t'} \subset S_{t,t'}$. Mais si $m_t \cap m_{t'} = \emptyset$ on ne sait pas si $m_{t'}$ comporte les personnes "nées du ménage m_t", celles qui y appartenaient à t étant décédées, ou si il s'agit d'un remplacement complet (les problèmes concret de ce type concernent les phénomènes de migrations plus que les naissances-décès habituels). On peut vérifier que la définition suivante :

$$\exists\, t_0 = t < t_1 < t_2 \quad < t_n = t' : \forall\, i\, (i = 0, \ldots, n-1) : m_{t_i} \mathrm{D}\, m_{t_{i+1}} \subset S_{t_i,\, t_{i+1}} \text{ et } m_{t_i} \mathrm{D}\, m_{t_{i+1}} \neq 0 \;,$$

capture correctement la notion de mouvement naturel. Reste à formaliser les autres "événements démographiques" que peuvent subir les ménages et les critères d'identification logique qu'on peut mettre en œuvre pour établir une filiation des identifiants. Ce n'est pas de la tarte. Prenons pour exemple un cas simple : l'individu $k \in m_{t-\varepsilon}$ se met en ménage avec l'individu $\ell \in m'_{t-\varepsilon}$. C'est la seule modification qui touche la population $m_{t-\varepsilon} \cup m'_{t-\varepsilon}$, de sorte qu'à $t+\varepsilon$ on a trois ménages :

$$\begin{aligned} m^1_{t+\varepsilon} &= m_{t-\varepsilon} - \{k\} \\ m^2_{t+\varepsilon} &= m_{t-\varepsilon} - \{\ell\} \\ m^3_{t+\varepsilon} &= \{k, \ell\}. \end{aligned}$$

On voit facilement que l'identification de ce qu'on peut appeler des "nouveaux ménages" est très arbitraire. On remarque par exemple $m^1_{t+\varepsilon}$ et $m^2_{t+\varepsilon}$ peuvent très bien être vides . Le concept de "chef de ménage" ou de "personne de référence " peut aider à ces problèmes d'identification, mais on sait bien à quel point leurs définitions sont arbitraires.

L'identification de logements est relativement plus facile. Ils sont généralement repérés par une adresse, un identifiant fiscal ou tout autre. Les notions de construction, d'achèvement et de destruction sont assez rigoureuses et correspondent au "mouvement naturel" de la population. Les seuls problèmes (hormis celui des habitations mobiles peut-être ?) est celui des regroupements et éclatements de logements. Ces événements sont cependant suffisamment rares et faciles à traiter statistiquement (voir § 8) pour ne pas gêner l'analyse longitudinale.

De façon générale, l'analyse longitudinale suppose une codification rigoureuse des événements démographiques survenant à la population d'études. On peut alors de façon non ambiguë et cohérente établir des règles de création, de suppression et de filiation des identifiants c'est-à-dire des unités statistiques de la population d'étude.

C'est possible pour les personnes et, dans une large mesure, pour les logements. On est très loin de ce pré-requis en ce qui concerne les ménages de sorte que c'est au moins un abus de langage (et au plus un non-sens) que de parler de "panel de ménages".

2 - Population fixe dans le temps

La population d'étude est constante (dans le temps) si $U_t = U_{t_0} = U_{t_f}$ pour tout t de T. On parle alors classiquement d'une cohorte. Cette situation résulte très généralement d'une convention qui définit cette cohorte à partir d'une population évolutive : ensemble des personnes qui ont subi le même événement au même

moment (c'est la définition classique des démographes), personnes survivantes à la date t_f , personnes figurant dans la même liste à la date t_0 (les "conscrits"), voire même étoiles d'un catalogue pour un panel à visées astronomiques.

Pour ce type de population, les problèmes sont relativement simplifiés. L'échantillonnage ne pose aucun problème particulier puisqu'on peut en principe travailler sur une base de sondage fixe et connue. Le recueil d'information, en revanche, n'a pas de solution simple et peut revêtir plusieurs aspects.

La méthode la plus rapide et la moins coûteuse est celle de l'enquête rétrospective. Son plus beau fleuron à l'Insee est sans doute, dans sa forme classique, l'enquête sur les familles associée aux recensements pour l'étude de la fécondité, de la nuptialité et de l'activité des femmes. Le principal problème qu'elle pose est celui des erreurs de mémoire (le "biais" de mortalité étant négligeable, semble-t-il). On pourra se référer par exemple à [J.C Deville (1972)].

Dans les méthodes d'observation suivie on peut distinguer les techniques d'observation longitudinales, qui sont des cas particulier de panels, comme les panels d'élèves suivis par les services de l'Education Nationale. Elles se caractérisent par le fait que l'échantillon est unique pour toute la période d'étude et que des données sont collectées pour l'ensemble des dates d'observation.

Les autres méthodes peuvent être qualifiées d'observation partielle. On y observe les données relatives à une unité échantillonnée *k* uniquement à certaines dates dépendant de *k*. Les méthodes les plus fréquemment utilisées sont celle des échantillons tournants (ou rotatifs) et celle des échantillons coordonnés, qui est plus générale.

Un échantillon rotatif résulte du partage d'un échantillon global *s* (exemple : les aires de l'enquête-emploi) en sous-échantillons s_a suivis sur une sous-période fixée T_a de T. Chaque s_a est en quelque sorte un panel sur la période T_a et l'échantillon rotatif peut être vu comme une famille de panels qui se chevauchent, c'est-à-dire relatifs à des périodes T_a qui se recouvrent. L'échantillon rotatif est conçu pour être extrapolable à chaque période d'observation $F \subset T$ et posséder les mêmes propriétés statistiques. Ceci implique (en première approximation et sous certaines conditions) que les sous-échantillons soient de même taille et observés le même nombre de fois.

Les schémas de rotation peuvent aller du très simple (enquête emploi annuelle classique) à l'assez compliqué (future enquête emploi en continu) si on veut obtenir une efficacité dans la mesure des évolutions à court terme et à moyen terme (trimestriel, mensuel ou annuel).

Dans un échantillonnage coordonné dans le temps, la période d'observation de chaque unité k est définie indépendamment de toute référence à l'appartenance à un sous échantillon. La méthode des numéros aléatoires est fréquemment utilisée dans les enquêtes auprès des entreprises et permet, par exemple, de faire entrer en observation une entreprise qui grossit en fonction de la taille qu'elle atteint. On se reportera à [Cotton, Hesse (1992)] et à [Hesse (1994)] pour plus de détails.

3 - Enquêtes longitudinales et enquêtes continues

Ça n'a quasiment rien à voir.

Dans tout ce papier (sauf ce paragraphe) nous supposons qu'il n'y a aucune ambiguïté sur la définition du temps et de la date d'observation. En particulier chaque vague d'enquête a lieu à une date bien déterminée $t \in F \subset T$. Quand on y regarde d'un peu plus près, ça change. Les différentes unités ne sont pas enquêtées au même moment exactement. On s'en tire en admettant que les variations possibles de date d'enquête n'induisent que des variations minimes sur les variables d'intérêt. On tolère donc une erreur de mesure $\left(\text{collecter } y_k^{t+e} \text{ au lieu de } y_k^{(t)}\right)$ qu'on juge négligeable. On peut aussi chercher, par appel à la mémoire, à obtenir à la date $t + \varepsilon$ la valeur de $y_k^{(t)}$. Ceci revient à tolérer une autre erreur de mesure liée au biais de mémoire et à estimer qu'elle est négligeable.

Dans certains cas, en fait quand des variations rapides des variables d'étude doivent être prises en compte, cette tolérance devient dangereuse. On sait par exemple, que la date de collecte de la défunte enquête-emploi trimestrielle est une des causes de l'instabilité des résultats qu'elle faisait apparaître. L'existence de phénomènes saisonniers instables (liés au climat : "la" vague de froid annuelle, la date des vendanges pour prendre un exemple célèbre [Le Roy-Ladurie (1967)]) peut perturber complètement la statistique. En particulier une collecte à date fixe peut rater complètement un phénomène saisonnier survenant plutôt tard ou, au contraire, de façon exceptionnellement précoce. Une façon de ne pas rater ces variations temporelles de survenue d'événements inévitables est de réaliser une opération d'enquête à collecte continue.

L'exemple le plus classique (du moins théoriquement car la réalisation pratique laisse beaucoup à désirer) est celui des enquêtes sur les budgets de famille. Dans ces enquêtes, l'échantillon est réparti de façon plus ou moins aléatoire sur la période d'étude. On observe donc $y_k^{(t_k)}$ où (k, t_k) est un aléatoire fixé par le plan de sondage. Le but est d'estimer des quantités de la forme $\sum_U \int_T y_k^{(t)}\, dt$ à partir des observations. Le caractère aléatoire des t_k fixés par le plan permet d'obtenir des

estimateurs sans biais à partir du moment où toute date t a, en un certain sens, une probabilité non nulle de faire l'objet d'une enquête.

Le fait de réaliser une enquête en continu n'exclut pas celui de recueillir des données longitudinales de type panel. C'est en principe, ce qui sera fait pour la future enquête-emploi. Et ce qui fait aussi une des difficultés, complètement sous-estimée, de cette opération.

4 - Populations renouvelées dans le temps et panels

Un panel est un échantillon adapté à l'étude d'une population qui se renouvelle au cours de la période d'étude T. En particulier il a la propriété d'être extrapolable à toute population U_t pour $t \in F$ (date d'observation) - *propriété transversale* - et de contenir toutes les données relatives à une unité k pour toutes dates d'observation $F_k = \{t ; t \in F \text{ et } k \in U_t\} = F \cap T_k$ où l'unité k appartient à la population d'étude - *propriété longitudinale* - .

Ces exigences semblent a priori beaucoup plus fortes que tout ce que nous avons envisagé jusqu'à maintenant. C'est tout à fait vrai en ce qui concerne la pratique de l'échantillonnage, de la collecte et de la gestion des données (y compris leur analyse). Sur le plan théorique, une petite astuce rend les choses beaucoup plus manipulables et permet de définir un cadre logique où tout s'organise assez bien.

Poursuivons la discussion commencée au paragraphe [1]. La période d'étude T est fixée, la population d'étude $U = \bigcup_{t \in T} U_t$ aussi. Un plan de sondage de type panel est donc une loi de probabilité p(s) sur l'ensemble des parties de U. Celle-ci permet de définir les probabilités d'inclusion π_k des individus dans le panel, ainsi que les probabilités d'inclusions doubles $\pi_{k\ell}$ permettant en principe de se livrer à des calculs et à des estimations de variance.

A chaque individu k de U est associée l'indicatrice $T_k = \{t \in T : k \in U_t\}$. C'est , comme on l'a dit, un intervalle ou une réunion d'intervalles. L'indicatrice d'inclusion est la famille de variables $E_k^{(t)} = 1$ si $k \in U_t$ et zéro sinon. Elle nous livre directement les statistiques à établir sur l'état de la population. L'effectif de celle-ci à l'époque t est $N_t = \sum_U E_k^{(t)}$ qui est estimé par $\hat{N}_t = \sum_U \frac{E_k^{(t)}}{\pi_k}$ si on utilise l'estimateur de Horvitz-Thompson. Les variables d'intérêts (ou auxiliaires) $y_k^{(t)}$ sont définies et collectables uniquement sur T_k, en principe. Complétons les de façon arbitraire sur $T - T_k$. On obtient ainsi une structure identique à celle des populations fixes dans le

temps. Dans notre formalisme celles-ci se caractériseraient par le fait que pour tout k, $E_k^{(t)}$ est constante et égale à 1. Une caractéristique fixe de k est une variable $x_k^{(t)}$ constante sur T_k (et sur T). Exemple : Lieu de naissance.

On peut dire aussi que $\left(U, y_k^{(t)}\right)$ est un processus aléatoire au sens mathématique du terme (la tribu sur U étant l'ensemble des parties). D'une certaine manière, contrairement au célèbre article de [G. KALTON (1993)], nous avons procédé à la suppression de la quatrième dimension.

L'intérêt de cette construction est le suivant. On sait que toutes les statistiques qui présentent un certain intérêt dans les problèmes de sondage peuvent être vues comme des totaux ou des fonctions de totaux. C'est le cas des effectifs, des moyennes, des ratios, des taux d'évolution, des corrélations entre variables (et donc des corrélations temporelles). Dans le cas des populations évolutives, on ne s'intéressera (sauf peut-être exception - le concours est ouvert) qu'à des "totaux vivants" c'est-à-dire de la forme $\sum_U y_k^{(t)} E_k^{(t)}$. Bien évidemment tous ces totaux sont estimables dès que l'estimateur de Horvitz-Thompson est disponible, c'est-à-dire que les π_k sont connus.

Donnons juste quelques exemples.

Le total de y à t est $\sum_U y_k^{(t)} E_k^{(t)} = \sum_U y_k^{(t)} = Y_t$. Il s'estime par $\hat{Y}_t = \sum_s y_k^{(t)} E_k^{(t)} / \pi_k = \sum_{s_t} \frac{y_k^{(t)}}{\pi_k}$ où $s_t = s \cap U_t$. On remarque que π_k ne dépend pas de t ce qui est la caractéristique d'un panel (en fait certaines variables auxiliaires, éventuellement datées, présentes dans la base de sondage peuvent déterminer le choix des π_k, comme dans tout problème d'échantillonnage. Nous ne développerons pas plus ces considérations dans ce papier). La moyenne des $y_k^{(t)}$ est le ratio $\frac{Y_t}{N_t} = \overline{Y}_t$ et s'estime par $\hat{\overline{Y}}_t = \frac{\hat{Y}_t}{\hat{N}_t}$.

L'évolution nette de la moyenne de y entre t et s est $\overline{Y}_t - \overline{Y}_s$ et s'estime par $\hat{\overline{Y}}_t - \hat{\overline{Y}}_s$.
L'évolution brute est l'évolution de la moyenne entre t et s de la population présente aux deux époques.
C'est le ratio $\frac{\Delta_{ts} Y}{N_{ts}}$ où $N_{ts} = \sum_U E_k^{(t)} E_k^{(s)}$ et $\Delta_{ts} Y = \sum_U \left(y_k^{(t)} - y_k^{(s)}\right) E_k^{(t)} E_k^{(s)}$ (alors que pour l'accroissement de y c'est $Y^t - Y^s = \sum_U \left(y_k^{(t)} E_k^{(t)} - y_k^{(s)} E_k^{(s)}\right)$) . Toutes ces

quantités sont des totaux ou des fonctions de totaux et s'estiment donc de façon naturelle et simple à partir de l'estimateur de Horvitz-Thompson. Le calcul et l'estimation de la variance de ces statistiques ne pose rigoureusement aucun problème nouveau.

Bref, tant qu'on se limite à l'estimateur de Horvitz-Thompson et que les données sont complètes (éventuellement imputées), on n'a aucun problème particulier avec les données de panel.

5 - Modalités d'exploitation d'un panel, nature et rôle de l'information auxiliaire et correction de la non-réponse

On a coutume de poser les deux assertions suivantes :

– un panel doit pouvoir être exploité transversalement,
– un panel doit s'exploiter à sa date d'échéance (ce qui pose problème si $t_f = \infty$!).

Ne faisons pas dans la dentelle : un panel doit pouvoir s'exploiter pour toute famille de date $F_0 \subset F$ à condition que $\sup_F t \leq \theta$ ou θ désigne la date actuelle (moins délai de préparation des données).

Les cas les plus habituels sont :

- $F_O = \{t^*\}$ *: où* $t^* = \sup\{t \in F \,;\, t \leq \theta\}$: exploitation transversale.
- $F_O = \{t_O, t^*\}$: Evolution depuis l'instant origine. C'est le cas des <u>indices</u> de type classique, Paasche ou Laspeyres.
- $F_O = \{t^{**}, t^*\}$ *ou* $t^{**} = \sup\{t \in F \,;\, t < t^*\}$: évolution récente.
- $F_O = \{t \in F \,;\, t \leq \theta\}$ ensemble des données connues, analyse longitudinale "complète" à la date actuelle.

Tant que les données sont complètes (non-réponse compensée par une imputation <u>définitive</u>) et qu'on utilise l'estimateur de Horvitz-Thompson, il n'y a aucun problème : toutes les statistiques sont estimables, l'estimation de variance peut se faire (au bémol des imputations près) et, surtout, il y a cohérence parfaite entre toutes les exploitations qu'on peut envisager.

L'ennui, c'est qu'on ne procédera jamais de cette manière.

En effet, le mode de correction de la non-réponse et l'amélioration de l'estimation par incorporation d'information auxiliaire va dépendre du choix de F_0 et de θ (car l'information externe disponible dépend de ce paramètre. On peut par exemple différer une exploitation transversale pour profiter de la disponibilité attendue d'une information auxiliaire - la pyramide des âges à t^* par exemple). Si on ne se pose pas de problèmes de cohérence des statistiques (et donc d'éventuelles révisions des résultats d'exploitations antérieures du panel), nous sommes ramenés à un problème standard : estimer des statistiques portant sur les variables $y^{(t)}$ pour $t \in F_0$ en utilisant, pour la correction des non-réponses et l'estimation, l'information auxiliaire disponible à savoir :

- *Les* $Z^{(t)}$ pour les t disponibles $(\leq \theta!)$
- *Les* $x_k^{(t)}$ pour $t \in F_0$
- *Les* $x_k^{(t)}$ *et les* $y_k^{(t)}$ pour $t \in F - F_0$ $(et\ t \leq \theta)$

En particulier les $y_k^{(t)}$ $t \notin F_0$, par exemple les valeurs antérieures de $y_k^{(t)}$ pour une analyse transversale, sont à considérer comme des variables auxiliaires utilisables dans la procédure d'estimation.

Examinons par exemple le cas de l'analyse transversale comparé à celui de l'analyse longitudinale complète à date actuelle.

On peut, conformément aux habitudes, décider de compenser la non-réponse complète par repondération. Celle-ci utilisera les données auxiliaires disponibles pour l'époque t^* mais éventuellement aussi un modèle de réponse qui aura pu être étalonné sur les périodes antérieures. Combiné avec l'information externe, on obtiendra des poids transversaux $w_k^{t^*}$ permettant l'analyse des données transversales, les estimations ponctuelles et les estimations de variance. La non-réponse partielle pourra être imputée grâce à toutes les variables auxiliaires disponibles, que celles-ci soient transversales où longitudinales. C'est le cas, par exemple, quand on utilise une imputation par ratio du style :

$$\hat{y}_k^{t^*} = y_k^{t^{**}} \left(\frac{\overline{y}^{t^*}}{\overline{y}^{t^{**}}} \right)_r$$

où le ratio est calculé sur les répondants communs aux époques t^* et t^{**}.

Ainsi, l'analyse transversale est de nature parfaitement classique. Elle se caractérise seulement par l'abondance (si le panel est ancien) et la nature de l'information auxiliaire disponible.

L'analyse longitudinale ne nécessite pas de nouvelles techniques. La non-réponse totale sera évidemment repondérée en fonction de l'information liée à la base où de l'information externe actuelle (attention, cependant, celle-ci doit être vue comme relative à la population U, et pas à la population U_t. Autrement dit tout calage "longitudinal" sur une donnée externe X_t relative à U_t s'écrira $\sum_s E_k^{(t)} x_k^t w_k = X_t$).

Le problème spécifique réside dans le fait que la non-réponse totale à une vague du panel doit être considérée comme une non-réponse partielle. Si on s'en tient aux habitudes, ce type de non-réponse est compensée par imputation, en utilisant toute l'information antérieure ou postérieure (interpolation à la date considérée).

Une constatation générale vient tempérer quelque peu cette façon de voir. Une grande partie de la non-réponse totale après la première vague est définitive (en tous cas on n'a pas observé de retour dans le champ des répondants à l'époque t^*!). C'est le phénomène d'érosion (in English attrition) d'un panel. Cette érosion est particulièrement forte après la première (et parfois aussi la seconde) vague.

Il paraît peu esthétique de procéder à des imputations massives d'unités présentes à seulement une ou deux époques de l'étude longitudinale. On peut procéder de la façon suivante. Supposons, pour fixer les idées, qu'on ne décide d'imputer la non-réponse totale qu'à partir de la troisième vague, la seconde étant en quelque sorte celle où on considère que la fidélisation au panel se stabilise. Les répondants au panel longitudinal seront alors, par convention, ceux de la seconde vague (échantillon r).

Il s'agit de ne pas jeter totalement l'information apportée par les répondants $\left(\text{échantillon } r_1 \text{ , } \; avec\; s \supset r_1 \supset r\right)$ de la première vague. Pour ce faire, on pourra fabriquer un jeu de pondérations longitudinales (relatives à l'échantillon r) compatible avec l'information apportée par r_1. On y parviendra en ajoutant aux équations de calage estimant les paramètres du modèle de non-réponse - qui sont de la forme $\sum_r w_k x_k = X$ - de nouvelles équations :

$$\sum_r w_k z_k = \sum_{r_1} w_k^{(1)} z_k = Z .$$

Là-dedans, Z désigne le vecteur des variables de la première vague dont on désire respecter l'information et $w_k^{(1)}$ le jeu des pondérations transversales adoptées pour l'exploitation de la première période. z_k peut contenir, par exemple, des indications relatives aux catégories d'activité à l'époque 1 si l'érosion se différencie surtout selon

ces variables. Cette technique se généralise sans problème technique particulier à la prise en compte de plusieurs périodes. Il faudra simplement prêter attention aux choix des statistiques sur lesquelles on décide de caler.

Conclusion provisoire :

L'exploitation des données collectées dans un panel se ramène à des techniques bien connues que ce soit en matière de correction de la non-réponse ou de choix d'estimateur. Le problème réside surtout dans les choix des contraintes conduisant à une certaine forme de calage ; ce choix peut s'avérer un peu délicat si on introduit des contraintes générales de cohérence entre les exploitations longitudinales et transversales.

6 - Quelques problèmes liés à l'échantillonnage

Nous avons ramené, de façon un peu formelle il est vrai, le problème de l'échantillonnage pour panel à celui de l'élaboration d'un plan de sondage p(s) sur la population d'étude $U = \bigcup_{t \in T} U_t$. Si celle-ci est connue à l'instant initial d'échantillonnage, il n'y a pas de problème (Exemple : Population des personnes de 18 ans et plus ayant vécu en France jusqu'à cet âge ; panel avec un horizon de 8 ans ; la base de sondage est un recensement réalisé l'année qui précède la mise en route du panel).

Généralement, malheureusement, ce n'est pas le cas. On utilise une base de sondage pour la population U_{t_0} avec un plan $\left(p_0(s)\,;\, s \subset U_{t_0}\right)$ qui permet de tirer l'échantillon longitudinal, celui qu'on utiliserait seul si on s'intéressait à une analyse par cohorte. Cet échantillon verra sa taille diminuer au cours des périodes successives par le jeu de la "mortalité" (= sortie de champ en général) et de l'érosion (qu'il faudrait arriver à distinguer le mieux possible même si ce n'est pas simple : un "mort" est aussi un non-répondant définitif !). Maintenant, les périodes successives $(t,\, t + \Delta t)$ seront, pour l'échantillonnage, traitées comme des strates avec une base de sondage spécifique $\beta_t^{\Delta t}$ (quand cela est possible !) . La limite de ce système est celui de l'enregistrement continu avec enrichissement continu de la base. Chaque unité sondable arrive dans la base à un instant spécifique. Elle peut être mise en réserve jusqu'au tirage (cas stratifié) ou être échantillonnée tout de suite. La seule technique utilisable alors est celle de l'échantillonnage Poissonnien où l'unité est incluse dans l'échantillon avec une probabilité π_k qui ne dépend que de ses caractéristiques propres et de l'information auxiliaire. Les pratiques qui consistent à rééchantillonner en fonction de l'érosion du panel (pour conserver un nombre de répondants fixe d'une vague à l'autre) n'ont aucune justification liée à la représentativité de

l'échantillonnage ; elles conduisent simplement à attribuer à la strate $\beta_t^{\Delta t}$ un poids d'extrapolation lié au taux d'érosion des vagues précédentes.

Remarque :

> Le fait de compléter le panel par un nouvel échantillon probabiliste dans $U_{t+\Delta t}$ est de nature différente et complique singulièrement les choses. Quand cet échantillon est exploité avec l'échantillon panel, on utilise généralement des pondérations qui supposent les deux échantillonnages indépendants (meilleur estimateur linéaire sans biais). En effet, généralement, les probabilités d'inclusions ne sont pas toujours calculables facilement. De fait si $\tilde{s} = s \cup s'$ on aura $\tilde{\pi}_k = P_r(k \in \tilde{s}) = P_r(k \in s) + P_r\left(k \in s' \middle| k \notin s\right)$.
>
> Généralement, si k vient effectivement de s, le second terme sera malaisé à récupérer, et inversement si k vient de s'.

7 - L'échantillonnage indirect

De fait, on échantillonne rarement les personnes (pour un panel ou pas !) à partir d'une base de sondage de personnes. Dans la pratique française du Panel Européen, l'échantillonnage est réalisé à partir de logements, qui permettent d'attraper les personnes, et, de façon intermédiaire, les ménages, qui sont un ensemble de personnes habitant un logement siège d'une résidence principale (sur l'articulation entre ces termes et une approche rigoureuse de ces définitions on pourra se reporter à [J-C.Deville (1988)], dans un ensemble de documents qui ne fut pas jugé digne d'entrer dans le sanctuaire de *Données Sociales*, où les données sont vraiment considérées comme données).

Ces logements sont échantillonnés dans l'échantillon-maître (EM) - ce qui permet une extrapolation à l'ensemble des logements recensés en mars 1990 - et dans la Base de Sondages des Logements Neufs (BSLN), qui est un panel des autorisations de construire décernées depuis 1987. Ce panel est lui-même échantillonné sur une base géographique qui a le bon goût, malgré diverses chicaïas administratives et la dérive des continents, d'être fixe dans le temps.

Autrement dit, le panel de logement qu'est l'échantillon-maître complété de la BSLN, permet d'entretenir un panel de personnes. Voyons comment.

La vague initiale sera échantillonnée à partir d'un échantillon habituel de logements. Chaque vague successive - théoriquement - doit être enrichie d'un échantillon de logements construits depuis la vague précédente. On construit ainsi un panel de logements (sous-panel de la base EM + BSLN !). Ce panel de logements a la

propriété d'être extrapolable transversalement à toute époque d'exploitation potentielle du panel de personnes. On obtient donc un échantillon extrapolable de personnes de la façon suivante :

– A t_0 (population longitudinale), l'échantillon est constitué de toutes les personnes du champ (condition d'âge éventuelle) trouvée dans les logements de l'échantillon de logements L_{t_0}. Ces personnes seront suivies jusqu'à l'horizon du panel, quel que soit le logement qu'elles occupent aux dates successives d'enquêtes.

– Pour toute date ultérieure t , soit L_t l'échantillon du panel de logements à l'époque t (échantillon initial L_{t_0} + logements neufs) ; on inclut dans le panel de personnes celles qui sont entrées dans le champ ("naissances", en pratique immigrés) depuis la date d'enquête antérieure.

Comme la majorité des personnes ne déménagent pas entre deux périodes d'enquêtes, cette méthode est assez économique.

Remarque :

En fait les "naissances" de bébés sont traitées un peu différemment. On utilise la particularité qu'ils ont d'avoir une mère, qui, éventuellement, peut avoir la chance de faire partie du panel.

8 - Echantillonnage indirect et "panels" de ménage

Ce qu'on vient d'analyser est un cas particulier d'échantillonnage indirect dont on va donner maintenant l'ébauche d'une théorie un peu plus générale et dont les applications sont multiples et même innombrables.

Une population U est échantillonnée selon un plan de sondage $(p(s)\,;\,s \subset U)$ autorisant à faire des statistiques grâce à des probabilités d'inclusion π_k et des techniques d'estimation comme cela a déjà été évoqué. On cherche à atteindre une population V d'individu courant i. Une matrice $A = \{a_{ki}\}$ de nombres positifs ou nuls relie ces deux populations. Pour que l'affaire marche bien, comme on le verra, il faut que la matrice A soit assez creuse, les a_{ki} non nuls étant rares.

Echantillonnant l'unité k de U, on enquêtera toutes les unités i de V telles que $a_{ki} > 0$. Les a_{ki} sont collectables auprès de l'unité k (et ne sont pas nécessairement connus dans toute la base de sondage). Lors de l'enquête auprès de l'unité i de V on collecte

également les a_{ki} positifs pour i fixé. Formellement, donc s_U est tiré dans U selon le plan p.

On en déduit un échantillon $s_V = \{i \in V \, ; \exists k \in s_U \text{ et } a_{ki} > 0\}$.

On collecte :
- les $a_{ki}>0$ pour $k \in s_U$
- les $a_{ki}>0$ pour $i \in s_V$

On suppose que pour tout i de V il existe au moins une unité k de U telle que $a_{ki}>0$ de façon à ce que toute la population V puisse être ainsi attrapée.

Donnons quelques exemples bien connus :

Exemple 1 :
U est la population des logements k, $V = \{i\}$ est un ensemble de personnes ; $a_{ki} = 1$ si M^r^ ou Mme i habite le logement k. Sinon $a_{ki} = 0$.

Exemple 2 :
$V = \{i\}$ peut aussi être un ensemble de ménages et alors $a_{ki} = 1$ si k est la résidence principale du ménage i (et vaut zéro sinon). Dans ce cas V est une partie de U puisqu'on identifie ménage et résidence principale.

Exemple 3 :
Sondage en grappes (c'est la généralisation de l'exemple 1).

Exemple 4 :
On admet que tout enfant de moins de 10 ans vit avec une personne majeure de plus de 18 ans inscrite sur la "Liste des Personnes Majeures". On tire un échantillon dans cette liste pour attraper des petits enfants. La relation est $a_{ki} = 1$ si l'enfant i habite avec la grande personne k. On notera que à chaque k, peuvent être associés plusieurs i et inversement.

Exemple 5 :
Un ensemble d'entreprises $V = \{i\}$ est possédé par des actionnaires k qu'on attrape parce qu'ils paient des impôts ; a_{ki} est le montant du capital de i possédé par k. On a ici un exemple de nature numérique particulièrement intéressant car on peut faire des transformations comme par exemple ne pas prendre en

compte les a_{ki} inférieurs à un certain seuil : $a'_{ki} = a_{ki}\ si > a$ et égal à 0 sinon ; ou aussi $a''_{ki} = 1$ si $a_{ki} > a$, etc.

Exemple 6 :
Entretien d'un panel de logements.
La démographie des logements est assez simple. Outre la phase prénatale (autorisation, mise en chantier, ...), nous sommes préoccupés par l'achèvement, la destruction et deux transformations assez simples : la fusion et l'éclatement (la recomposition peut également s'envisager sans trop poser de problème). Dans tous les cas, si nous nous intéressons à la population "avant" et "après" l'événement, la matrice A est définie par $a_{ki} = 1$ si le logement k "avant" participe au logement i après. Pour ce qui concerne la gestion des identifiants une règle arbitraire d'héritage en cas de fusion, ou de "déclinaison" en cas d'éclatement, peut être appliquée selon ce qu'on entend par logement identique "après" et "avant".

Le procédé peut s'itérer pour attraper une troisième population W, d'individu courant j, liée à V par une matrice $B = (b_{ij})$, d'éléments positifs ou (le plus souvent) nuls. Sur ce principe, on peut aussi faire des enquêtes en deux phases en échantillonnant dans l'échantillon s_V. Bref la procédure est assez riche de possibilités. Avant de théoriser là-dessus, annonçons le dernier exemple.

9 - Comment faire des statistiques sur les ménages dans un panel, suivi de, comment enrichir un panel sans trop se fatiguer

La principale application de l'échantillonnage indirect est la possibilité d'attraper des ménages grâce aux individus. Dans la pratique française c'est automatiquement réalisé lors de la première vague puisque nous démarrons sur un échantillon en grappes. Dès la vague suivante, on doit se poser le problème à cause de la démographie bizarroïde et mal connue des ménages. De plus, les individus de ces ménages "transversaux" (comme on dit !) sont eux-mêmes objet d'enquête car l'information qu'ils peuvent apporter s'avère non seulement non inutile mais surtout pas chère à collecter.

Les matrices qui relient les "personnes-panels" aux ménages et aux individus de ces ménages sont les suivantes :

– $a_{ki} = 1$ si la personne-panel k est dans le ménage i

– $b_{ij} = 1$ si la personne j (de U_t) appartient au ménage i

– $c_{kj} = \sum_i a_{ki}\, b_{ij}\ (C = AB)$ si la personne j appartient au ménage de la personne-panel k.

On pourra remarquer que dans B on peut introduire des caractéristiques particulières des personnes permettant une exploitation (cf. paragraphe 10) sur des populations particulières. On peut aussi enrichir le panel sans trop se fatiguer en chaînant le procédé et en panélisant les individus ainsi attrapés. Ceci demande un certain doigté pour une exploitation longitudinale. Pour faire du transversal, en revanche, comme on le verra, tout va bien (de même, ce qui est assez chouette, que pour faire des statistiques d'accroissement net !).

La façon dont on traite les bébés est une autre application, fort utile. L'accroissement de population $U_{t+\Delta t} - U_t$ comporte deux parties : les immigrants (hors champ à t mais vivants) et les bébés (nés entre t et t + Δt). On a vu comment on pouvait attraper les immigrants par la technique du panel de logements. On pourrait faire la même chose avec les bébés mais il serait un peu gênant (quoique ?) de sélectionner seulement le bébé d'une famille qui vient de s'installer dans un logement panel, ou de suivre les personnes du ménage rien qu'à cause de ce foutu bébé. L'alternative consiste à capturer les bébés par leurs parents. Donc, dans la matrice A, on aura $a_{ki} = 1$ si l'individu "panel" (ou pas !) k est parent du bébé i. Pour des raisons d'incertitude génétique et de tradition des démographes, on aura en fait $a_{ki} = 1$ si l'individu "panel" (ou pas) k est la mère du bébé i. Le bébé en question peut alors être panélisé avec un poids qui va apparaître dès le paragraphe suivant.

10 - Pondération et échantillonnage indirect. La méthode de partage des poids

– Pondération : On s'intéresse au total $\sum_V y_i = Y$ d'une variable y de la population V. Si on note 1_V le vecteur avec des 1 pour chaque indice i et $y = (y_i)$ on peut écrire $Y = 1_V' \cdot y$ (produit scalaire). Posons $a_{.i} = \sum_{k \in U} a_{ki}$. On a l'identité $Y = \sum_{k,i} \frac{a_{ki} y_i}{a_{.i}}$. On suppose - c'est essentiel n'est-ce pas - que $a_{.i}$ a été collecté à l'enquête. Dans le cas des personnes des ménages transversaux, $a_{.i}$ est tout simplement le nombre de personnes panélisables présentes dans le ménage dont i fait partie (et donc le nombre de personnes du ménage tout court si le panel est sans restriction, d'âge par exemple).

La variable $z_k = \sum_{i \in V} a_{ki} \frac{y_i}{a_{.i}}$ est donc définie pour tout k de U et mesurée pour tout k de s_U. On a, bien évidemment, $\sum_U z_k = \sum_V y_i$. Notons qu'on peut écrire $Z = A\, diag\left(A' 1_U\right)^{-1} y = \tilde{A} y$ si on aime les notations matricielles. Si on dispose de pondérations w_k sur les individus panels (par exemple les poids de l'estimateur de Horvitz-Thompson), on dispose aussi d'un estimateur de

$$Z = \sum_k z_k = 1_U' z = \left(1_U' . A\right) diag\left(A' 1_U\right)^{-1} y = 1_V' . y = Y \text{ par } \hat{Z} = \hat{Y} = \sum_{s_U} w_k z_k = w' . z$$

si on note w le vecteur des w_k.

Sous cette forme il est donc clair qu'on obtient un estimateur de $Z = Y$ sans biais si les poids sont sans biais et dont on sait estimer la variance si on sait le faire pour les poids w_k.

On peut, de même, estimer toute fonction de totaux (estimateur par substitution) ainsi que la variance de cette statistique (linéarisation). Bref tous nos problèmes sont résolus.

On peut aller plus loin, et profiter d'une information auxiliaire que ce soit au niveau de la population U (ce qui va de soi) mais aussi au niveau de la population V, et bien entendu, des deux simultanément. Voyons d'abord la forme opérationnelle (fichier de dépouillement) de cette méthode. Transformons l'estimateur $\hat{Z}$.

$$\hat{Z} = \sum_{s_U} w_k z_k \sum_{s_U} w_k \sum_{i \in V} a_{ki} \frac{y_i}{a_{.i}} = \sum_{s_V} y_i \sum_{s_U} \frac{w_k a_{ki}}{a_{.i}} = \sum_{s_V} w_i^* y_i$$

Matriciellement :

$$\hat{Z} = w' \tilde{A} y = w^{*'} y \text{ avec } w^* = w' \tilde{A}.$$

On a $w_i^* = \frac{1}{a_{.i}} \sum_k w_k a_{ki}$ où, évidemment, la somme ne porte que sur les $w_k a_{ki}$ non nuls, c'est-à-dire sur les individus k "rattachés" à i par la positivité des a_{ki}. Dans le cas des "individus transversaux" du panel, la somme est celle de tous les poids des "personnes panels" appartenant au ménage de i, $a_{.i}$ est le nombre de personnes (du champ) de ce ménage. D'où le terme de partage des poids [Ernst (1989)] donné à cette méthode longtemps considérée comme empirique. On remarquera que si le ménage de V a la même composition que le ménage de U qui "pointe" sur lui, les individus ont pour poids la moyenne des w_k (ce qui ne change rien si ces poids

sont égaux). En revanche les personnes attrapées par une seule personne-panel se voient contraintes à partager le poids de cette dernière.

Indiquons enfin comment on peut profiter de l'information auxiliaire présente au niveau de la population V. Nous admettons que celle-ci se présente sous la forme d'un vecteur X de totaux connus sur V. Le respect de l'information nous demande donc de satisfaire l'équation de calage :

$\sum_{s_V} w_i x_i = X = w'\tilde{A}\underline{x} = \sum_{s_U} w_k \left(\sum_i \tilde{a}_{ki} x_i \right)$ *où* $\underline{x}$ est la matrice qui empile des x_i.

On est donc ramené à un problème standard de calage, si on le désire. Le partage des poids sera alors un partage des poids calés sur s_U. On pourrait concevoir un calage direct des poids w_i^* eux-mêmes. La question de la cohérence avec les calages sur U doit alors s'étudier, mais, en ce qui me concerne, plutôt un autre jour.

14 - Conclusion

Que ce soit dans les aspects liés à l'échantillonnage ou dans les aspects liés à l'estimation, les enquêtes par panels ne posent essentiellement pas de problèmes nouveaux. La difficulté réside surtout dans la mise en œuvre des procédures "classiques", dans le fait d'identifier la nature exacte des problèmes. En particulier, certaines difficultés surviennent quand on cherche à mettre en cohérence des exploitations relatives à des ensembles de dates différents (transversal et longitudinal par exemple).

Enfin, la technique d'échantillonnage indirect, qu'on pratique déjà sans le savoir dans les enquêtes ponctuelles, devient un outil essentiel dans la statistique de panels. Il faut savoir identifier les situations où elle s'avère utile et la mettre en œuvre à bon escient. Elle permet toutes les exploitations transversales imaginables et peut même permettre, si on n'est pas trop difficile, d'entretenir un panel sur une durée indéfinie.

Références et bibliographie :

BINDER D.A, "Longitudinal surveys : why are these surveys different from all other surveys ?" IASS/IAOS Satellite Meeting on Longitudonal Surveys, Jerusalem, August 27-31, (1997)

CHAMBAZ C, SAUNIER J.M, VALDELIEVRE H, "Méthodologie du panel européen de ménages : exploitation des données de la vague 2 du fichier français", Insee, Direction des Statistiques Démographiques et Sociales, document de travail F 9715 (1997)

COTTON F, HESSE C, "Méthodes d'échantillonnage pour l'enquête annuelle d'entreprises, Actes des JMS de 1991, Insee Méthodes, vol. 29-30-31, (1992)

DEVILLE J.C, *Structure des familles : résultats de l'enquête de 1962,* Collections de l'Insee, Vol. D, 13-14, (1972)

DEVILLE J.C, "Peut-on croire aux enquêtes ?" dans *Construire les données sociales,* Collections de l'Insee, Vol. M 128, pp. 15-22, (1988)

ERNST L.R, "Weighting issues for longitudinal and family estimates", dans *Panels Surveys*, edited by KASPRZYK D, DUNCAN G, KALTON K, and SINGH M.P Wiley (1989)

HESSE C, "Tirage, rotation, retirage d'un panel stratifié de taille fixe : la méthode panastra", Insee, Direction des Statistiques d'Entreprises, (1994)

HOLT D, SKINNER C.J "Components of change in repeated surveys", International Statistical Review, Vol. 57, pp. 1-18, (1989)

KALTON G, CITRO C.F, "Enquêtes par panel : ajout d'une quatrième dimension" Techniques d'Enquête, Vol. 19, pp. 217-227, (1993)

LAVALLEE P, Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids" Techniques d'Enquête, Vol. 21, pp. 27-35, (1995)

LAVALLEE P, "Représentativité et pondération dans les enquêtes longitudinales", Université de Caen (1996)

LEROY-LADURIE E, *Histoire du climat depuis l'an 1000,* Flammarion (1997)

BINDER D.A, "Longitudinal surveys : why are these surveys different from all

QUELLES MESURES PRENDRE POUR LIMITER LES EFFETS DE L'ATTRITION D'UN PANEL LORS DE LA COLLECTE? L'exemple du panel européen de ménages

Dominique Ansieau

1- Introduction

L'attrition est un fléau pour les enquêtes par panel. Ses causes sont multiples. La lassitude des unités enquêtées et les refus qui en découlent sont des éléments qui contribuent à l'attrition. Mais une mauvaise gestion de l'enquête peut elle-même devenir une source importante d'usure de l'échantillon. Avec le temps elle peut même prendre le pas sur l'usure due aux refus de collaborer. Or la gestion d'un panel de ménages, en intervenant tout à la fois au niveau du ménage et de l'individu, est particulièrement complexe. La mise en place des outils adéquats pour combattre l'attrition y devient primordiale.

Le Panel européen de ménages est un exemple de panel de ménages. Il s'agit d'une enquête multidomaine réalisée, à l'initiative d'Eurostat, dans l'ensemble des pays de l'Union Européenne (exception faite de la Suède) depuis 1994. Le principal objectif de l'enquête est la description de la situation d'activité de chaque individu adulte du ménage, de l'évolution dans le temps de cette situation ainsi que des revenus perçus par le ménage et par chaque individu qui le compose. Des calendriers d'activité au mois par mois y sont renseignés et les revenus y sont détaillés au niveau individuel en une cinquantaine de rubriques. D'autres domaines tels la formation, les relations sociales, la santé, ou quelques éléments de biographie sont également abordés. L'ensemble des personnes adultes du ménage sont interrogées individuellement sur la base d'un Questionnaire Individuel. Les informations concernant le logement, les revenus de placement par exemple, quant à eux sont abordés au sein d'un Questionnaire Ménage.

Les ménages interrogés lors de la première vague d'enquête de 1994 ont ensuite été ré-interrogés chaque année. Le Panel Européen des Ménages s'arrêtera vraisemblablement en 1999 après avoir enquêté auprès des mêmes ménages six ans de suite.

Reprendre contact avec les mêmes ménages chaque année peut paraître simple comme règle de suivi d'un panel de ménages. Mais cela cache une multiplicité de situations parfois complexe à gérer. La première difficulté est dans le fait que l'entité ménage sur laquelle reposent bon nombre d'enquêtes « classiques » devient floue et instable dans la durée. La définition habituelle d'un ménage au sens Insee correspond à : « L'ensemble des personnes habitant un même logement ». Cette définition statique devient inapplicable dans le temps. Les individus vivant ensemble dans un même logement à un instant donné, peuvent changer de logement ou encore ne plus vivre ensemble à un autre moment.

Les procédures de suivi adoptées dans une enquête telle que l'enquête Emploi privilégient la notion de logement. On y ré-interroge les occupants d'un même logement d'une année sur l'autre que ses occupants soient les mêmes ou pas. Les Panels de ménages quant à eux privilégient l'individu. Ainsi, bien que l'ensemble du ménage continue d'être interrogé au cours d'une vague d'enquête, les règles de suivi sont établies au niveau individuel.

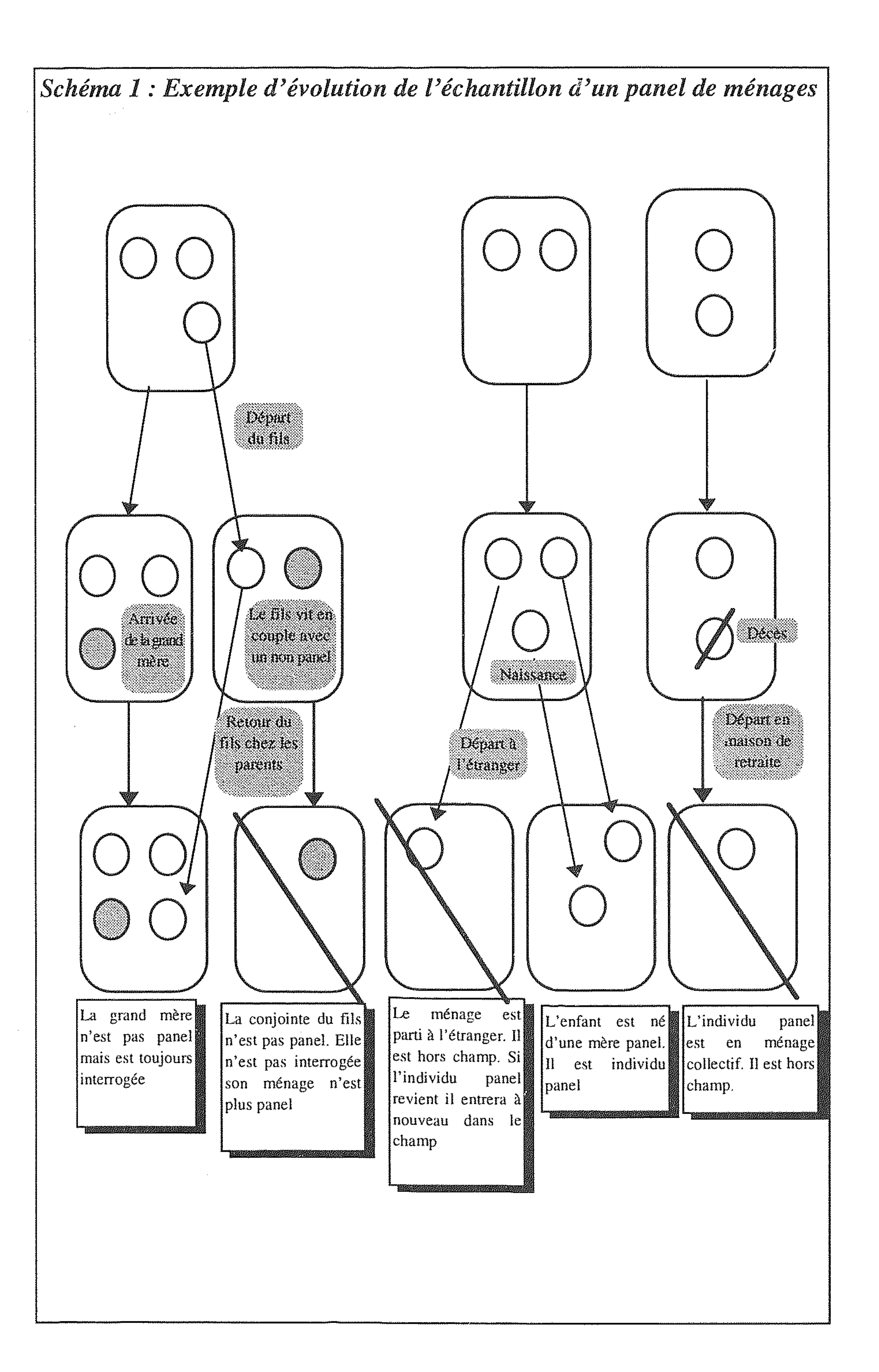
2- Quelques définitions

Les individus interrogés lors de la première vague sont qualifiés *d'individus panels*. Les enfants d'une mère elle-même panel et nés depuis la première enquête sont également définis comme *individus panels*. Seuls ces *individus panels* sont par la suite suivis systématiquement et interrogés de vague d'enquête en vague d'enquête. Cette définition propre au Panel Européen des Ménages peut varier d'un panel à l'autre sur quelques points de détail. Mais le principe général reste le même.

La notion de ménage quant à elle change. Elle est dans le cadre d'un panel de ménages étroitement liée à l'individu panel et s'affranchit du logement initial. Le *ménage panel* est, lors d'une vague d'enquête donnée, l'ensemble des personnes qui vivent dans un même logement avec au moins un individu panel.

Enfin les individus appartenant à un ménage panel sans qu'ils soient eux mêmes panels sont interrogés au même titre que les individus panels tant qu'ils continuent à vivre dans un ménage panel. En revanche dès qu'ils quittent un tel ménage ils ne sont plus suivis. Ils sont souvent appelés individus « non-panels ».

Schéma 1 : Exemple d'évolution de l'échantillon d'un panel de ménages



Le schéma n°1 montre comment l'échantillon d'un panel de ménages peut dès la deuxième vague d'enquête se modifier. Les règles de suivi peuvent différer à la marge d'une enquête à l'autre. Mais les modifications dans la structure de l'échantillon en cours même de collecte sont communes à tous les panels de ménages. Les modifications de l'échantillon ne se limitent d'ailleurs pas aux créations et disparitions successives de ménages. La répartition géographique des ménages à enquêter évolue également. Chaque création d'un nouveau ménage entraîne en général un changement d'adresse. Mais de nombreux ménages qui ne changent pas pour autant de structure sont également confrontés à des déménagements. Dans le Panel Européen des ménages environ un ménage sur sept est concerné au cours de chaque vague d'enquête par un changement d'adresse, que celui-ci s'accompagne ou non d'un changement de composition.

Les ménages qui ne changent ni de composition ni de localisation sont bien entendu les plus faciles à retrouver et à ré-interroger. Les refus de collaborer sont la principale source d'attrition qui touche cette partie de l'échantillon. En revanche les ménages qui changent d'adresse ou qui « éclatent » en plusieurs nouveaux ménages sont en général plus difficiles à contacter. Ne pas gérer correctement ces cas parfois complexes revient à accepter une perte importante de l'échantillon d'une vague à l'autre.

3- Limiter les refus dans un panel

Les ménages qui refusent de collaborer à l'enquête sont de moins en moins nombreux au cours des vagues successives. La plupart d'entre eux refusent dès la première ou la deuxième vague d'enquête. Toutefois pour limiter ces refus complets ou partiels de répondre, il convient de mettre le ménage en confiance.

3.1 Des mesures actives

Conserver d'une année sur l'autre, et autant que faire ce peut, le même enquêteur pour chaque ménage enquêté, contribue à créer une relation qui favorise la participation du ménage à l'enquête. L'enquêteur étant déjà connu du ménage ce dernier répond plus en confiance. Ainsi sur les 5 740 ménages de la vague 2 qui ont été interrogés par le même enquêteur qu'en vague 1, le taux de refus s'élève à 6,6 %. Alors que si on ne retient que les 744 ménages qui n'ont pas changé d'adresse mais qui ont été interrogés par un enquêteur différent ce taux s'élève à 9,8 %.

Renvoyer chaque année le même enquêteur auprès des mêmes ménages peut également influencer la qualité des données collectées. Mais là les effets sont plus difficiles à évaluer. Le ménage interrogé peut donner plus facilement des données qu'il juge sensibles (par exemple les revenus) à partir de la deuxième vague

d'enquête. En revanche l'enquêteur connaissant peut-être trop bien la situation du ménage est parfois tenté de ne pas poser certaines questions ou de ne pas les poser correctement.

La pratique de maintenir le même enquêteur pour le même ménage est très bonne pour diminuer l'attrition. En revanche les effets sur la qualité des données collectées est plus discutable. Toutefois les effets sur l'attrition sont en général jugés suffisamment importants pour la maintenir.

Reprendre contact avec les ménages qui refusent est bien sûr très important. Et dans le cadre d'un panel cette relance doit pouvoir être plus personnalisée que dans une enquête classique. On connaît en effet quelques informations sur le ménage. Cette relance peut également chercher à déterminer les raisons du refus de coopérer. En particulier les refus liés à la personnalité de l'enquêteur ou à un événement familial peuvent plus facilement se rattraper.

Pour terminer une dernière mesure visant à limiter les effets des refus de collaborer sur l'attrition à moyen terme consiste à ré-interroger à chaque vague les ménages qui ont refusé lors de l'enquête précédente. Les ménages qui refusent deux fois de suite sont exclus de ce processus. Cette mesure n'est pas exempte d'inconvénient. En particulier elle génère des trous dans les séries chronologiques collectées. Cependant on peut ainsi « rattraper » d'une année sur l'autre une partie des ménages qui ont refusé. Les refus fermes sont rarement récupérés de cette manière. Mais une partie des refus sont plus liés aux circonstances du moment (deuil dans la famille, perte d'emploi...) qu'à un véritable rejet de l'enquête.

Cette procédure est employée dans le cadre du Panel Européen des ménages depuis la deuxième vague d'enquête. Et environ 15 % des ménages qui avaient refusé de répondre en 1995 ont finalement accepté de le faire l'année suivante.

3.2 Des mesures préventives

L'éventail des mesures prises pour impliquer les ménages, ou au moins créer un climat de confiance, peuvent être très variées. Leur mise en place dépend des moyens qui y sont consacrés tant en terme budgétaire qu'en terme humain. La remise de résultats tirés de l'enquête aux ménages lors des différents contacts est une pratique courante. Relativement peu coûteuse à mettre en place cette opération nécessite cependant de présenter ces résultats sous une forme attrayante et lisible par la majorité des personnes enquêtées. L'équipe du Panel Canadien présente ainsi ces résultats sous forme d'un jeu de questions réponses. Dans le cadre du Panel Européen, des résultats ont été communiqués à l'occasion de chacun des contacts pris avec le ménage. L'effet réel sur l'attrition est très difficile à mesurer. Il est de

toute façon plus préventif qu'actif. La présentation de résultats permet de montrer « l'utilité » de l'enquête et donc de la contribution qu'y apporte le ménage.

Offrir un cadeau au ménage pour le remercier de sa collaboration est également assez fréquent. Toutefois cette pratique est plus onéreuse que la précédente. Dans le cadre du Panel Européen un cadeau est remis au ménage après chaque vague d'enquête terrain depuis 1995. L'effet de ces cadeaux sur le taux de réponse des ménages est très difficile à évaluer, même s'il s'avère que dans certains cas ils ont permis à l'enquêteur de réaliser l'enquête. Le choix du cadeau est délicat. Celui-ci doit en effet être apprécié de tous, ne pas être en décalage avec l'image de l'Institut, et surtout être différent chaque année. L'option américaine d'offrir un cadeau adapté à chaque ménage et choisi par l'enquêteur qui gère son propre budget « cadeau » n'est pas encore à l'ordre du jour en Europe.

Enfin une relation épistolaire régulière et individualisée peut être entretenue avec les ménages de l'échantillon. Là aussi l'ampleur de ces relations dépend des moyens qui y sont consacrés. Pour le Panel Européen des Ménages qui se déroule sur le terrain au cours du dernier trimestre de chaque année, à l'issue de chaque vague d'enquête un courrier remerciant le ménage de sa collaboration et lui présentant nos voeux pour l'année à venir est systématiquement envoyé. Mais d'autres équipes ayant en charge la gestion d'un panel vont, par exemple, jusqu'à envoyer des cartes pour les anniversaires des individus de l'échantillon.

Les moyens consacrés à l'opération sont une limite à la mise en place de ces pratiques. Mais de toute façon le cumul de celles-ci rend leur efficacité de plus en plus marginale. Le nécessaire équilibre à trouver entre le coût et l'efficacité de ces mesures dépend également de la durée de l'opération. Ainsi un Panel où l'échantillon est renouvelé tous les trois ans consacrera certainement moins de moyens à réduire l'attrition qu'un panel qui interrogera les mêmes ménages pendant dix ou vingt ans.

4- Limiter l'attrition liée aux mouvements de personnes : organisation de la collecte

Les règles de suivi d'un panel de ménages sont établies au niveau individuel. Ainsi l'échantillon va fluctuer en fonction du mouvement des individus qui composent chaque ménage. Les ménages vont ainsi changer d'adresse, ou encore voir leur composition se modifier.

4.1 Le mouvement des personnes

Le mouvement des personnes a ainsi des conséquences à la fois sur la structure de l'échantillon de ménages et sur sa localisation géographique. Ce sont les mouvements des individus panels qui importent. Cependant il est nécessaire d'être également attentif au mouvement des non-panels pour éviter la création de faux nouveaux individus.

Les mouvements des individus au sein d'un ménage peuvent se résumer aux catégories suivantes :

☞ Les individus qui ne bougent pas.

☞ Les individus qui entrent dans la composition du ménage.

☞ Les individus nés depuis la dernière enquête. Cette catégorie est un cas particulier de la précédente.

☞ Les individus qui sortent du ménage.

☞ Les individus qui décèdent. Comme pour les naissances, cette catégorie est un cas particulier de la précédente.

Ces mouvements d'individus se jugent par rapport au *ménage panel.* Cette proposition qui semble être une évidence n'est cependant pas toujours facile à mettre en œuvre. Elle va à l'encontre des réflexes pris en associant habituellement les notions de ménage et de logement.

En particulier si tous les individus panels d'un même ménage partent ensemble pour résider à une nouvelle adresse, ces individus ne seront pas considérés comme sortants. Dans ce cas c'est le ménage qui bouge. Les mouvements de chacun des individus panels entrent alors dans la catégorie « individus qui ne bougent pas ».

De même le mouvement des autres individus se juge également par rapport au ménage panel. Pour illustrer notre propos prenons un exemple :

> Soit un individu panel JEAN qui depuis la deuxième vague d'enquête vit avec un individu non-panel IRENE. En tant qu'individu non-panel IRENE ne sera suivie que si elle est accompagnée d'un individu panel.
>
> Dès la troisième vague JEAN quitte IRENE et rejoint ses parents MARCEL et JULIE qui sont également non-panels.

L'enquêteur qui prendra contact avec le ménage à l'adresse de la vague 2 va donc rencontrer IRENE qui lui apprendra que JEAN est parti et est retourné dans le ménage de ses parents. Or en terme de mouvement individuel on a une vision différente des choses.

Le ménage panel est défini par JEAN. Les mouvements sont donc les suivants :

JEAN ne bouge pas.

IRENE sort du ménage panel de JEAN.

MARCEL et IRENE entrent dans le ménage panel de JEAN.

Ainsi un individu panel ne quitte réellement un ménage que si par ailleurs au moins un autre individu panel ne bouge pas. Une fois ce mouvement repéré il faut aussi déterminer les conséquences en terme de suivi. En particulier il faut identifier si, à la suite de ce mouvement, l'individu sera toujours dans le champ de l'enquête.

Pour cela il sera nécessaire de répartir ces sorties d'individus en fonction de leur point d'arrivée (ménage ordinaire en France, ménage collectif, étranger etc.). Cette répartition pourra être variable d'un panel à l'autre en fonction du champ de l'enquête.

Identifier clairement le décès de l'individu par rapport aux autres sorties est très important pour plusieurs raisons. D'abord il s'agit a fortiori d'une sortie de champ, la disparition de l'individu qui s'ensuit ne fait donc pas l'objet d'un redressement. L'autre conséquence du décès de l'individu est qu'il ne risque plus de revenir dans le champ de l'enquête et ne nécessite donc pas de procédure de suivi particulière.

La gestion des entrées d'individus dans la composition d'un ménage est en général plus simple à gérer. Cependant il faut dans ce dernier cas être attentif à deux cas particuliers. Le premier concerne les naissances qui notamment dans la gestion du Panel Européen sont clairement identifiées. Elles peuvent en effet en fonction des définitions retenues pour l'individu panel avoir des conséquences sur l'échantillon et son suivi. Pour le Panel Européen les enfants nés de mère panel sont eux-mêmes panels. Il est donc primordial de repérer les entrées de type naissance ainsi que de repérer qui est la mère de l'enfant.

L'autre point délicat à propos des entrées d'individus dans la composition du ménage concerne le retour d'un individu au sein d'un ménage où il a déjà été enquêté. Il est nécessaire de repérer ces retours et éviter de créer de faux nouveaux individus.

Pour mener à bien cette tâche l'enquêteur doit donc au moins disposer de la composition du ménage à la dernière enquête pour pouvoir repérer les sorties mais aussi les entrées. Ce repérage ne peut évidemment se faire qu'en différentiel par rapport à une composition de référence. Mais pour pouvoir repérer plus facilement les retours d'individus on a également choisi de mettre à disposition de l'enquêteur l'historique de la composition du ménage depuis le début de l'opération.

Pour cela un document « permanent », le Tableau Permanent de Composition du Ménage a été créé. Celui-ci comporte la composition du ménage lors de la première vague d'enquête. Il est remis à l'enquêteur à chacune des vagues d'enquête successives pour être remis à jour. Cette façon de procéder permet à l'enquêteur d'avoir une liste de l'ensemble des personnes qui ont appartenu au ménage depuis le lancement de l'enquête. Il dispose ainsi des éléments nécessaires au repérage des éventuels retours d'individus dans le ménage.

4.2 Les conséquences au niveau ménage

Les règles de suivi d'un panel de ménage sont établies au niveau de l'individu. Toutefois les opérations de collecte sont organisées autour du concept de ménage. Les conséquences sur le niveau ménage des mouvements des individus vont donc avoir des effets sur l'organisation de la collecte.

On peut se trouver ainsi confronté à :

☞ Des ménages qui ne bougent pas. C'est le cas le plus fréquent.

☞ Des déménagements : Tous les individus panels d'un même ménage changent d'adresse tout en continuant de vivre ensemble.

☞ Des éclatements de ménages : Au moins deux individus panels vivant au sein du même ménage se séparent et forment désormais deux ou plus nouveaux ménages.

☞ Des fusions de ménages : Il s'agit d'un cas particulier des deux situations évoquées ci-dessus, où les individus panels concernés rejoignent un ménage panel déjà existant.

☞ Des disparitions de ménages : Le cas le plus fréquent de disparition de ménage est le décès de l'unique individu panel le composant. La fusion de ménages entraîne aussi généralement la disparition d'un des ménages concernés.

4.2.1 Les déménagements

L'une des principales difficultés liées à la gestion d'un déménagement est la recherche d'adresse. Si celle-ci n'aboutit pas le ménage est perdu pour le panel. C'est un des points clés de réussite d'un Panel. Et cependant il n'existe pas de recettes gagnantes à tous les coups. L'efficacité des outils utilisés dépend de l'insertion du ménage dans son environnement, de son intérêt pour l'enquête mais aussi de l'implication de l'enquêteur dans la démarche.

Pour faciliter la recherche des nouvelles adresses plusieurs outils peuvent être mis en place :

☞ Le contact intervague : Ce procédé consiste à prendre contact avec le ménage entre deux vagues d'enquête pour repérer à l'avance les changements possibles. Plusieurs méthodes peuvent être envisagées. Les plus fréquentes restent le contact postal ou le contact téléphonique.

Entre chaque vague d'enquête terrain du Panel Européen, au moins un contact a été pris au printemps avec les ménages. En 1995 et 1996 ce contact se déroulait sous forme d'une enquête téléphonique et ne concernait qu'une partie de l'échantillon. Depuis ce contact se déroule par voie postale et touche l'ensemble des ménages de l'échantillon.

☞ Les adresses relais : Ce système classique dans les enquêtes de type Panel consiste à obtenir auprès du ménage panel enquêté les coordonnées d'un autre ménage où l'enquêteur pourrait obtenir la nouvelle adresse du ménage panel en cas de déménagement.

Les contraintes fixées par la CNIL à l'utilisation de ces adresses relais dans le cadre du Panel Européen des Ménages en ont certainement limité les effets. Mais il s'agit d'un outil efficace d'aide à la recherche d'adresse.

☞ La rémunération des adresses retrouvées : Il ne s'agit pas à proprement parler d'une mesure qui facilite la recherche d'adresse. Cependant elle vise à motiver les enquêteurs à procéder à une recherche qui autrement ne lui rapporte rien, notamment si le ménage a quitté son secteur d'enquête. Ce point est d'autant plus important que même si certaines adresses sont obtenues facilement d'autres nécessitent que l'enquêteur dépense beaucoup d'énergie dans sa recherche sans garantie de résultat.

☞ Faire suivre le courrier ou ne pas le faire suivre ? Lors d'un contact postal avec le ménage, et en particulier lors de l'envoi de la Lettre avis prévenant le ménage du prochain passage de l'enquêteur il faut déterminer si on fait suivre le courrier ou pas.

Lors du contact préliminaire à l'enquête il est plus opportun de faire suivre le courrier. Ainsi le ménage qui a déménagé devrait tout de même recevoir la lettre qui lui est destinée. Il peut ainsi réagir et recontacter l'organisme responsable de la collecte. Lors des autres contacts la solution à retenir dépend de l'objectif. Ne pas faire suivre le courrier revient à gérer les retours et ainsi à repérer les ménages qui sont partis. Faire suivre le courrier privilégie en revanche le contact avec le ménage.

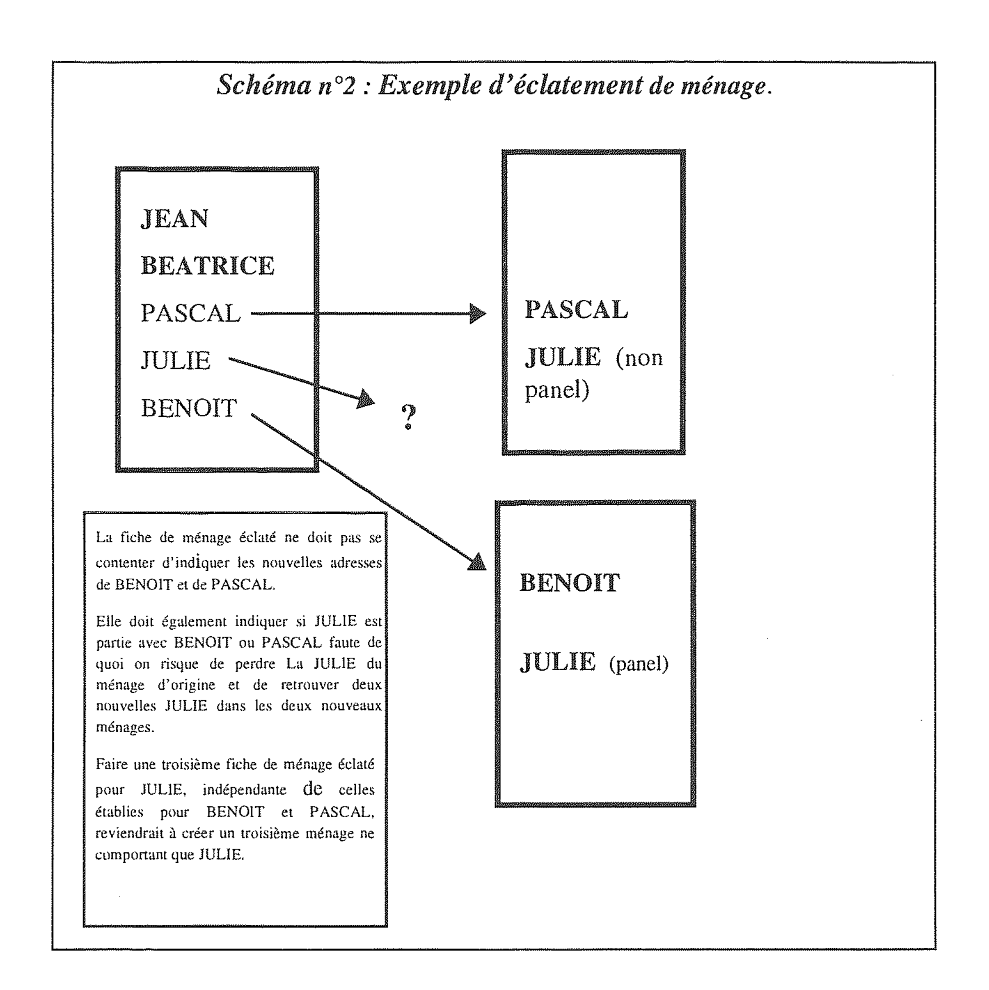
Le Panel Européen des Ménages est, à l'image de la plupart des enquêtes ménages de l'Insee, géré localement par les Directions régionales de collecte et coordonné centralement par une équipe nationale. La gestion des déménagements dans le cadre d'un panel s'en trouve compliquée. Elle peut en effet entraîner des transferts de dossiers d'un établissement régional à l'autre. Ces transferts doivent être sécurisés. Pour s'assurer que ceux-ci se passent dans de bonnes conditions un système d'avis d'envoi et d'accusé de réception a été mis en place pour le Panel Européen. Il est coordonné par l'équipe nationale. Ce système est lourd à gérer mais il permet de repérer rapidement les dysfonctionnements et de pouvoir y remédier.

4.2.2 Les éclatements de ménages

L'éclatement d'un ménage panel se produit dès que deux au moins des individus panels le composant se séparent pour former deux nouveaux ménages panels. Par définition l'éclatement de ménage se double d'un changement d'adresse pour au moins un des nouveaux ménages ainsi créés. Une partie de la gestion de ces ménages éclatés est donc commune à celle des déménagements.

Cependant il ne suffit pas, pour bien gérer ces cas particuliers, de repérer le départ d'un individu panel et de retrouver sa nouvelle adresse. Il est en effet absolument nécessaire de connaître de façon précise quels sont les individus (panels ou non) qui accompagnent l'individu panel qui quitte le ménage. Ne pas procéder de cette manière reviendrait soit à oublier des individus (voir schéma n°2 ci-dessous) soit à créer à tort des ménages éclatés supplémentaires.

Un mauvais repérage des individus qui ont quitté le ménage peut entraîner la perte de la trace des individus concernés, même si ceux-ci sont bien enquêtés dans le cadre du panel. Enquêter un même individu deux années de suite mais sans repérer qu'il s'agit du même revient à générer deux individus indépendants et à perdre tout l'intérêt longitudinal des données ainsi collectées.



4.2.3 La fusion de ménages

En général l'entrée d'un nouvel individu dans la composition d'un ménage panel n'a pas de répercussions sur la gestion de la collecte proprement dite. Il existe cependant une exception notable lorsque l'individu entrant est lui-même un individu panel provenant d'un autre ménage.

La fusion de ménages est le cas le plus délicat à résoudre. On dispose en général de très peu d'information pour la repérer. On considère également à tort qu'il s'agit d'un cas rare. Or à partir de la troisième vague d'enquête elle peut correspondre au retour au sein du giron familial d'un de ses membres qui l'a quitté au cours d'une vague d'enquête précédente, elle devient alors plus fréquente. Par ailleurs le mode d'échantillonnage choisi peut augmenter la probabilité de rencontrer de telles fusions. Ainsi le tirage de l'échantillon du Panel Européen dans l'échantillon maître

contribue à avoir des ménages qui sont proches géographiquement les uns des autres. La probabilité qu'il existe des relations entre ces ménages en augmente d'autant.

La mise en place du Tableau Permanent de Composition du Ménage (voir § 4.1) permet de résoudre relativement simplement la plupart des retours d'individus panels dans un ménage panel auquel ils ont déjà appartenu depuis le début de l'enquête. L'enquêteur peut, au vu de l'état civil de l'individu entrant, repérer qu'il existe déjà dans la liste.

L'entrée dans un ménage panel d'un individu panel qui n'y a jamais appartenu est plus problématique. Aucun outil fiable n'existe pour repérer ces situations. C'est finalement l'enquêteur qui est le mieux placé sur le terrain pour identifier ces cas qui restent heureusement rares. Certains panels nationaux utilisent les fichiers de population tenus dans leur pays pour repérer les entrants qui correspondent à des individus déjà enquêtés mais perdus depuis. Cette procédure est inapplicable en France. Quant à l'appariement d'un fichier d'individus perdus avec un fichier de nouveaux individus, il est difficilement envisageable. Compte tenu des différences orthographiques le prénom est pratiquement inutilisable comme clé d'appariement. Il ne reste alors que le sexe, le mois et l'année de naissance. Dans ces conditions le résultat de l'appariement ne peut être que très approximatif.

4.2.4 Une nécessité de gestion : la rapidité

L'ensemble de ces mouvements et leurs conséquences sur les opérations de collecte doivent souvent être gérés en continu pendant ces mêmes opérations. Même les changements prévus dès le contact intervague doivent être vérifiés sur le terrain.

Cette gestion simultanée à la collecte nécessite la mise en place d'une intendance conséquente. Un déménagement aboutit souvent à un changement d'enquêteur voire à un changement de région d'enquête. Il faut ainsi assurer le transfert du dossier d'un enquêteur à l'autre.

L'organisation de la collecte du Panel Européen des Ménages est basée sur la territorialité. Les différentes Directions régionales gèrent ainsi l'opération sur leur territoire. Les transferts de dossiers d'un enquêteur à l'autre sont donc gérés pour le Panel Européen à trois niveaux. Si le ménage reste dans le secteur d'enquête de l'enquêteur alors l'enquêteur s'en charge, il n'y a en fait pas de transfert. Si le ménage continue de résider dans la même région la Direction Régionale qui a en charge le dossier gère le transfert d'un enquêteur à l'autre. Dans les autres cas l'équipe nationale gère le transfert du dossier d'une Direction régionale à l'autre, l'attribution du dossier au bon enquêteur restant du ressort de la Direction Régionale.

Ces tâches d'intendance doivent à la fois être sécurisées et rapides dans leur mise en œuvre. Sécurisées parce qu'il faut éviter d'égarer les dossiers concernés, et bien sûr assurer la confidentialité pendant le transfert. Mais il faut aussi que ces transferts se fassent rapidement. Les opérations sont en effet limitées dans le temps. Il est important que le dossier transféré parvienne au nouvel enquêteur dans un délai qui lui permette de procéder à l'entretien.

Dans le cas d'un éclatement de ménage l'opération se complique puisqu'il y a création d'un nouveau ménage. Cette création doit s'accompagner de l'attribution d'un nouveau numéro de ménage. Pour éviter la création de numéros en double cette attribution ne peut se faire qu'au niveau de l'équipe nationale qui devient un point de passage obligé. La procédure de transfert s'en trouve légèrement alourdie (voir aussi § 5.1 sur les identifiants).

Les impératifs de sécurité, forcément consommateurs de temps, entrent en conflit avec la nécessaire rapidité de la procédure mise en œuvre. Il convient là aussi de trouver un juste équilibre.

5- L'attrition et l'organisation des données

Ne pas perdre d'individus pendant la phase de collecte ne signifie pas pour autant que la partie est gagnée. Une mauvaise organisation des données peut également être source d'attrition. La qualité des variables de gestion et la sécurisation des identifiants des différentes unités sont primordiaux pour le traitement d'un panel.

5.1 Les identifiants individuels

Dans les enquêtes ménages habituellement réalisées à l'Insee, l'identification des unités statistiques est centrée sur le concept de ménage. Ainsi le numéro de chaque individu est déduit de celui de son ménage en y adjoignant un numéro d'ordre qui correspond au numéro d'apparition dans la composition du ménage.

Une telle pratique pour un Panel de ménages devient dangereuse. En effet l'individu n'est plus attaché à un unique ménage au cours du temps. Baser son repérage sur un numéro de ménage qu'il quittera peut-être bientôt risque d'entraîner des erreurs. En fait cette identification « hiérarchique » subsiste dans l'enquête Panel européen. Mais elle est essentiellement utilisée au sein d'une même vague d'enquête. Un autre numéro, tout à fait indépendant du précédent a ainsi été créé. Ce numéro est conservé par l'individu tout au long de sa « vie panel ».

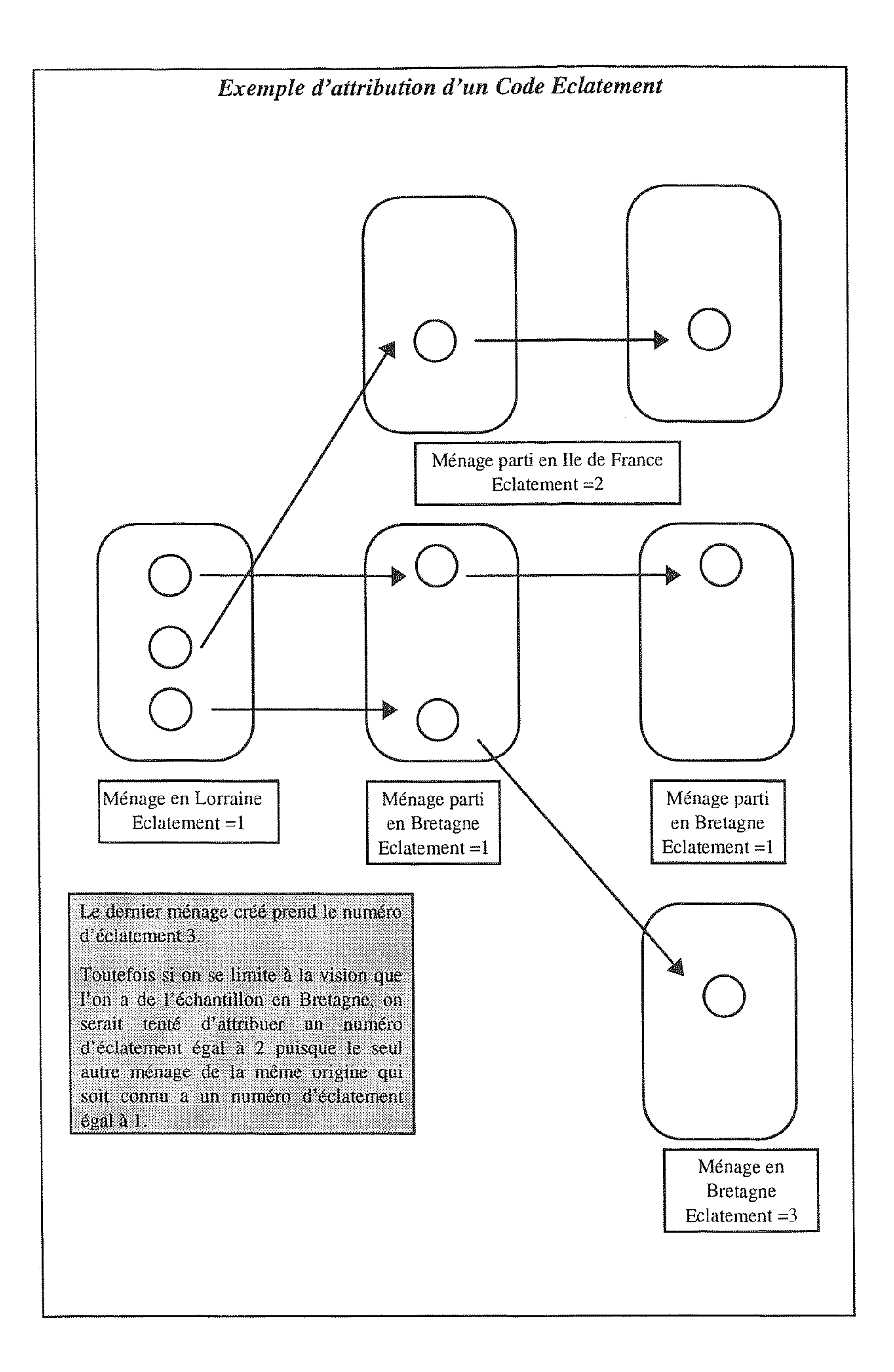
L'affectation d'un tel identifiant doit être là aussi gérée avec beaucoup de précaution. Il est en effet primordial de ne pas affecter le même identifiant à deux

individus différents. Dans le cadre de l'enquête Panel Européen on a utilisé un listing d'étiquettes. L'attribution du numéro individuel se fait alors en collant l'étiquette correspondante sur le document concerné.

5.2 L'identification du ménage

Bien que le ménage devienne une entité instable dans le temps, nous avons conservé dans le Panel Européen une filiation entre les numéros de ménages des vagues d'enquête successives. Deux raisons principales expliquent cette décision. La première tient au fait que conserver l'identifiant d'origine permet également de conserver l'accès aux informations de la base de sondage. La seconde tient au fait qu'une grande partie des ménages restent cependant stables d'une vague d'enquête à l'autre.

Enfin il peut sembler intéressant de conserver un lien entre les nouveaux ménages créés suite à un éclatement et leur ménage d'origine. Pour ce faire les nouveaux ménages ainsi constitués ont repris le numéro de leur ménage d'origine auquel on a adjoint un code d'éclatement. Créé en séquentiel pour le Panel Européen, ce code prend des valeurs comprises entre 1 et 9, la valeur 1 étant réservée aux ménages d'origine. Bien que son principe d'attribution soit simple, celle-ci ne peut être faite qu'au niveau central. La mobilité des ménages d'un échelon géographique à l'autre ne permet pas aux intervenants qui n'ont qu'une vision partielle de l'échantillon (c'est notamment le cas des enquêteurs et des Directions régionales qui gèrent l'enquête sur leur territoire) de l'attribuer sans risque d'erreur.


Exemple d'attribution d'un Code Eclatement
Ménage parti en Ile de France
Eclatement =2
Ménage en Lorraine
Eclatement =1
Ménage parti
en Bretagne
Eclatement =1
Ménage parti
en Bretagne
Eclatement =1
Le dernier ménage créé prend le numéro d'éclatement 3.
Toutefois si on se limite à la vision que l'on a de l'échantillon en Bretagne, on serait tenté d'attribuer un numéro d'éclatement égal à 2 puisque le seul autre ménage de la même origine qui soit connu a un numéro d'éclatement égal à 1.
Ménage en
Bretagne
Eclatement =3

L'attribution de ces numéros d'identification doit se faire avec beaucoup d'attention. Toute erreur dans ces affectations entraîne une perte des unités concernées en terme de suivi longitudinal et peut annihiler tous les efforts réalisés lors de la phase de collecte.

5.3 Les variables de gestion

Les variables de gestion sont nombreuses dans un panel. Elles permettent notamment :

☞ D'indiquer le statut de réponse de chaque ménage de l'échantillon (répondant refus, non joignable etc.)

☞ D'indiquer au sein de chaque ménage le statut de réponse de chaque individu. Un individu appartenant à un ménage qui a accepté l'enquête peut pour sa part refuser de répondre à son questionnaire individuel. Et les jeunes de moins de 17 ans ne sont pour leur part pas concernés par un tel questionnaire.

☞ D'indiquer le type de mouvement de chaque individu au sein du ménage (entrant, stable, sortant).

☞ De préciser si l'individu est panel ou non (important pour la mise en application des règles de suivi).

La qualité de ces variables est primordiale pour pouvoir procéder à un redressement dans de bonnes conditions. Mais elle l'est également pour l'organisation de la collecte elle-même.

En particulier la préparation de l'échantillon de ménages à enquêter lors d'une vague d'enquête dépend de son statut de réponse aux vagues précédentes. Il en va de même au niveau individu. Le type de questionnaire individuel qui lui est posé ne sera pas identique selon qu'il a déjà renseigné un questionnaire à la vague d'enquête précédente ou pas.

Ces variables de gestion ne sont pas toujours considérées comme importantes par les enquêteurs sur le terrain. Et leur qualité au cours des premières vagues d'enquête du Panel Européen était assez mauvaise. Ainsi, par exemple, de nombreux ménages étaient censés avoir refusé selon ces codes de gestion alors qu'ils avaient effectivement répondu à l'enquête. La difficulté de gérer à la fois un statut de réponse au niveau ménage puis au niveau individu a certainement perturbé le bon remplissage de ces variables. Ainsi un refus partiel d'un individu au sein du ménage

s'est souvent traduit par un refus de l'ensemble du ménage au travers des codes de gestion.

Il en va de même à propos des variables permettant de repérer les mouvements des individus au sein du ménage. En particulier le classement des entrées et des sorties ne s'est pas toujours fait dans les bonnes catégories. C'est notamment les cas pour plusieurs naissances qui ont été classées en « entrée classique ».

L'effet de ces erreurs d'affectation n'est pas forcément catastrophique s'il est repéré suffisamment tôt. Mais redresser ces informations demande un travail long et fastidieux. Le contrôle de qualité sur ces variables est non seulement indispensable mais doit par ailleurs être réalisé rapidement.

5.4 L'apport de CAPI

L'apport de CAPI (Collecte Assistée Par Informatique) dans la mise au point d'une enquête de type panel de ménages est très fort dans la constitution du questionnaire. CAPI permet notamment d'intégrer dans le questionnaire, et sans risque d'affectation, des informations collectées au cours des vagues antérieures en vue de les valider ou de repérer des changements de situations.

Mais dans le domaine qui nous intéresse ici CAPI permet surtout d'apporter une plus grande sécurité sur les identifiants. Contrairement aux enquêtes « papier » ils ne sont plus reportés manuellement d'un questionnaire à l'autre.

CAPI permet également d'avoir une grande cohérence entre les variables de gestion et le résultat de la collecte. Si l'enquêteur signale un refus de collaborer il ne peut pas en même temps continuer l'entretien.

Enfin les tests intégrés au questionnaire permettent de contrôler directement sur le terrain la qualité des variables de gestion qui ne sont pas directement prises en main par le système.

CALCUL DES PONDÉRATIONS DANS LE PANEL EUROPÉEN DE MÉNAGES

Christine Chambaz et Nadine Legendre

Introduction

Le panel européen de ménages a été lancé en 1994 dans les douze pays composant alors l'Union européenne. Son objectif principal est d'étudier la dynamique d'emploi et de revenus des *personnes*. En France, comme pour toutes les enquêtes auprès des ménages réalisées par l'Insee, l'échantillon initial est composé de logements. Ce sont d'abord les ménages qui sont contactés, même si les individus constituent la cible de l'enquête. L'enquête comporte d'ailleurs deux questionnaires, l'un adressé au ménage (sur sa composition, ses conditions de logement...), l'autre aux individus de 17 ans et plus (sur les revenus, l'emploi, la santé, les relations sociales,...).

Contrairement à ce que son nom laisserait supposer, le panel européen est donc *un panel d'individus*. Les règles d'évolution de l'échantillon ont effectivement été fixées en référence aux individus. A la base de ces règles, deux grands principes définissent deux catégories d'individus au regard de l'enquête :

- les individus composant les ménages tiré l'année 1 constituent l'échantillon de base du panel. Ils seront suivis annuellement tant que le panel durera, même s'ils déménagent. Ils forment la population des *« individus panel »*. Par définition, tous les individus des ménages répondant en vague 1 du panel sont des individus panel. Parmi eux, se trouvent des personnes ayant effectivement répondu au questionnaire individuel, mais aussi des individus n'y ayant pas répondu, soit qu'ils ont refusé, soit qu'ils étaient trop jeunes. Par convention, les enfants nés au cours des vagues suivantes de mère individu panel sont également des individus panel.

- les autres adultes des ménages dans lesquels se trouve au moins un individu panel sont interrogés, mais seulement aussi longtemps qu'ils restent avec cet individu panel (adulte ou enfant). Ils constituent la population des *« individus non-panel »*. Par définition, cette population n'existe qu'à partir de la vague 2 du panel, et devrait être de contours mouvants au fil des vagues à venir. Sa prise en compte permet, pour l'analyse transversale, de pallier en partie l'appauvrissement de l'échantillon consécutif aux sorties d'individus panel du

champ de l'enquête. Elle doit en principe assurer la représentativité instantanée de la population des ménages enquêtés[1].

Le panel européen des ménages n'est pas en fait un vrai panel, mais un suivi de cohorte. Il n'y a en effet pas de renouvellement de l'échantillon (voir Deville, 1998).

Compte tenu de l'existence de ces deux catégories de population répondant à l'enquête, deux types de pondérations sont donc calculées :

- une pondération longitudinale, propre aux individus panels ayant répondu en vague 1[2] , et répondant en vague 2. Cette pondération doit permettre l'étude des évolutions individuelles. Elle dérive des poids de base des individus, définis comme les poids initiaux corrigés de l'évolution de l'échantillon[3]. Pour une période donnée *(t,t+n)*, on peut calculer autant de pondérations longitudinales qu'on définit de sous-périodes de suivi. On peut par exemple, sur la période (1994,1996), décider de suivre les individus ayant répondu à la fois en 1994, en 1995 et en 1996 ; on peut aussi préférer regarder le parcours de ceux qui ont répondu en 1994 et 1996, ou en 1995 et 1996. Au sein d'un même ménage, il peut y avoir partage des poids de base : tous les individus panel répondants reçoivent alors le même poids longitudinal. En vague 2, cela n'a aucun intérêt : les poids longitudinaux correspondent aux poids de base. En revanche, cette opération sur les poids de base pourra ultérieurement permettre d'attribuer un poids longitudinal à des individus non-panel mais ayant répondu à plusieurs enquêtes successives.

- une pondération transversale, à utiliser pour l'analyse en coupe des réponses individuelles. Cette pondération est commune à tous les adultes répondants d'un ménage, qu'ils soient individus panel ou non.

1. A l'immigration près, sauf à considérer que les immigrants récents intègrent tous des ménages préexistant à leur arrivée. Compte tenu de la faiblesse des flux migratoires, cette restriction pourra, sur le court terme, être négligée.
2. Par commodité, et pour assurer la cohérence avec les hypothèses retenues lors du redressement de la vague 1 (non réponse individuelle au sein des ménages répondants considérée comme négligeable, et donc non redressée), tous les individus panel sont réputés avoir répondu en vague 1.
3. non-réponse, mais aussi déformation de la population des ménages. Théoriquement, seule la non-réponse devrait être corrigée pour le calcul de ces poids. Nous verrons cependant que le souci d'assurer une cohérence lors de la diffusion de résultats provenant d'une part d'analyses transversale, d'autre part d'analyses longitudinales, nous a conduits à recaler ces poids sur la structure de la population observée par ailleurs en 1995 (au moment de la réalisation de la vague 2).

Le calcul de ces pondérations a été effectué en accord avec la méthode recommandée par Eurostat, selon le schéma suivant[4] :

(i) identification des individus panel et de leur statut de réponse ;

(ii) analyse de la non-réponse individuelle ;

(iii) estimation de probabilités de non-réponse sur des groupes homogènes ;

(iv) correction de la non-réponse : calcul des poids de base en vague t (Vt) par correction des poids de base en vague (t-1) (Vt-1), en fonction de probabilités de non-réponse estimées sur des groupes homogènes ;

(v) calage de ces poids de base sur la structure par âge et sexe de la population en t, estimée à partir de l'enquête Emploi ;

(vi) calcul des poids longitudinaux et transversaux, avec et sans étape de calage.

Plusieurs modèles de non-réponse ont été testés. L'objet de cette étude est de présenter l'incidence du choix du modèle sur la distribution des poids et sur l'estimation de quelques indicateurs.

Dans une première partie, nous présentons une analyse des caractéristiques principales des individus panel adultes (17 ans et plus) non-répondants. Nous proposons ensuite, pour la vague 2, une correction de la non-réponse par groupes homogènes (CNRGH) à partir de deux modèles. L'un utilise les seules caractéristiques socio-démographiques des individus, l'autre intègre des variables de comportement, tel le déménagement consécutif à un éclatement du ménage-source.

Nous décrivons ensuite les distributions des poids longitudinaux obtenus selon les deux modèles, et des poids transversaux dérivés, calculés selon la méthode de partage des poids. Nous comparons enfin les valeurs de quelques indicateurs estimés à l'aide de l'un ou l'autre jeu de pondérations, sur la population totale, puis sur les sous-populations les plus mobiles (jeunes, personnes seules...). Les analyses présentées ici reposent essentiellement sur le redressement de la vague 2 du panel. Des éléments sur la correction de la non-réponse en vague 3 figurent en annexe 2

4. Une description plus complète du schéma théorique pourra être trouvée dans **EUROSTAT** (1995).

1. Evolution de l'échantillon entre les vagues 1, 2 et 3

1.1 L'attrition diminue entre les vagues 2 et 3

En vague 1 (V1), 7 344 ménages avaient été interrogés, soit environ 76 % des ménages appartenant au champ de l'enquête. Ils comprenaient 18 9916 individus, dont 14 524 adultes susceptibles de répondre au questionnaire individuel. Parmi ces derniers, 193 avaient refusé de répondre. La non-réponse individuelle au sein des ménages répondants était donc faible (1,3 %).

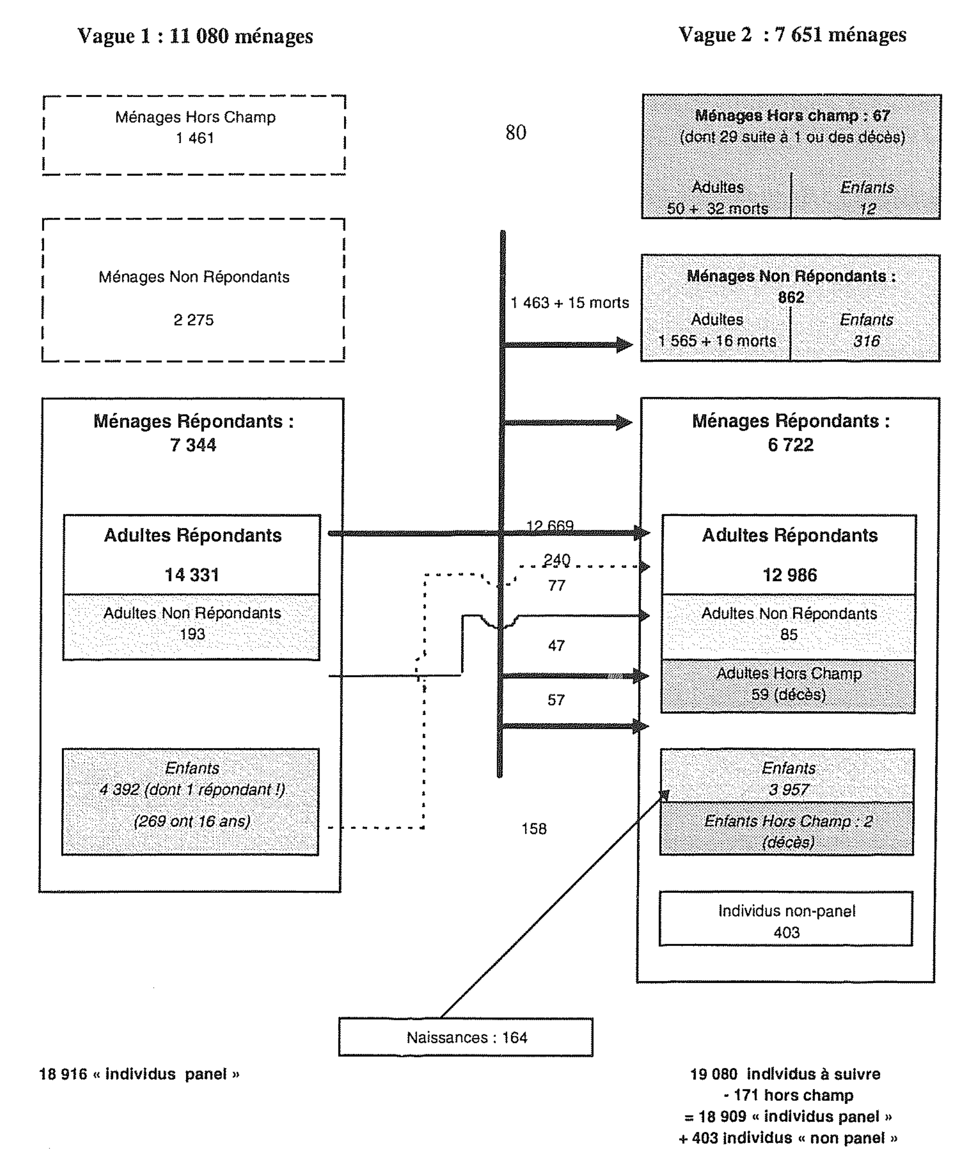
En vague 2 (V2), les individus des ménages non-répondants en V1 sont considérés hors champ. Seuls les 18 916 individus « panel » des 7 344 ménages *ayant répondu en V1*, et eux seuls, sont censés être réinterrogés. 171 sont cependant sortis du champ de l'enquête, par décès, déménagement en ménage collectif ou départ à l'étranger. A contrario, 164 nouveau-nés sont apparus. Au total, ce sont donc 18 909 individus panels à suivre, dont 14 636 adultes de 17 ans ou plus, susceptibles de répondre au questionnaire individuel. Parmi ces adultes, 268 avaient moins de 17 ans en V1.

Tous les individus à interroger ne se trouvent pas dans le ménage qui était le leur en V1. Un certain nombre de ménages ont en effet éclaté, par départ d'un de leurs membres ; il y a ainsi eu création de 307 nouveaux ménages. Le nombre de ménages à contacter pour réaliser les entretiens individuels s'est donc accru de 4,2 %, passant à 7 651 [voir schéma 1].

Les taux de réponse sont relativement bons, meilleurs qu'en V1 : 6 722 ménages ont répondu à l'enquête, soit 88,7 % des 7 584 ménages entrant dans le champ de l'enquête. Au niveau individuel, le taux d'acceptation de l'enquête est le même : 88,8 %, avec 12 986 adultes panel ayant accepté l'interview. Parmi les ménages répondants, 85 individus panel ont refusé de répondre (0,7 %, moins qu'en V1) ; 40 % d'entre eux avaient déjà refusé en V1. Les jeunes de 17 ans, nouvellement entrés dans le groupe adulte des individus panel sont nombreux à accepter l'enquête lorsque leur ménage avait répondu en V1 (89,3 %).

Le taux d'attrition, défini comme l'écart à 1 du rapport entre le nombre d'individus panel répondants en V2 et le nombre d'individus panel répondants en V1 est de 9,4 %.

Schéma 1 :Évolution de l'échantillon entre les vagues 1 et 2 - Schéma simplifié (devenir et origine des adultes répondants)

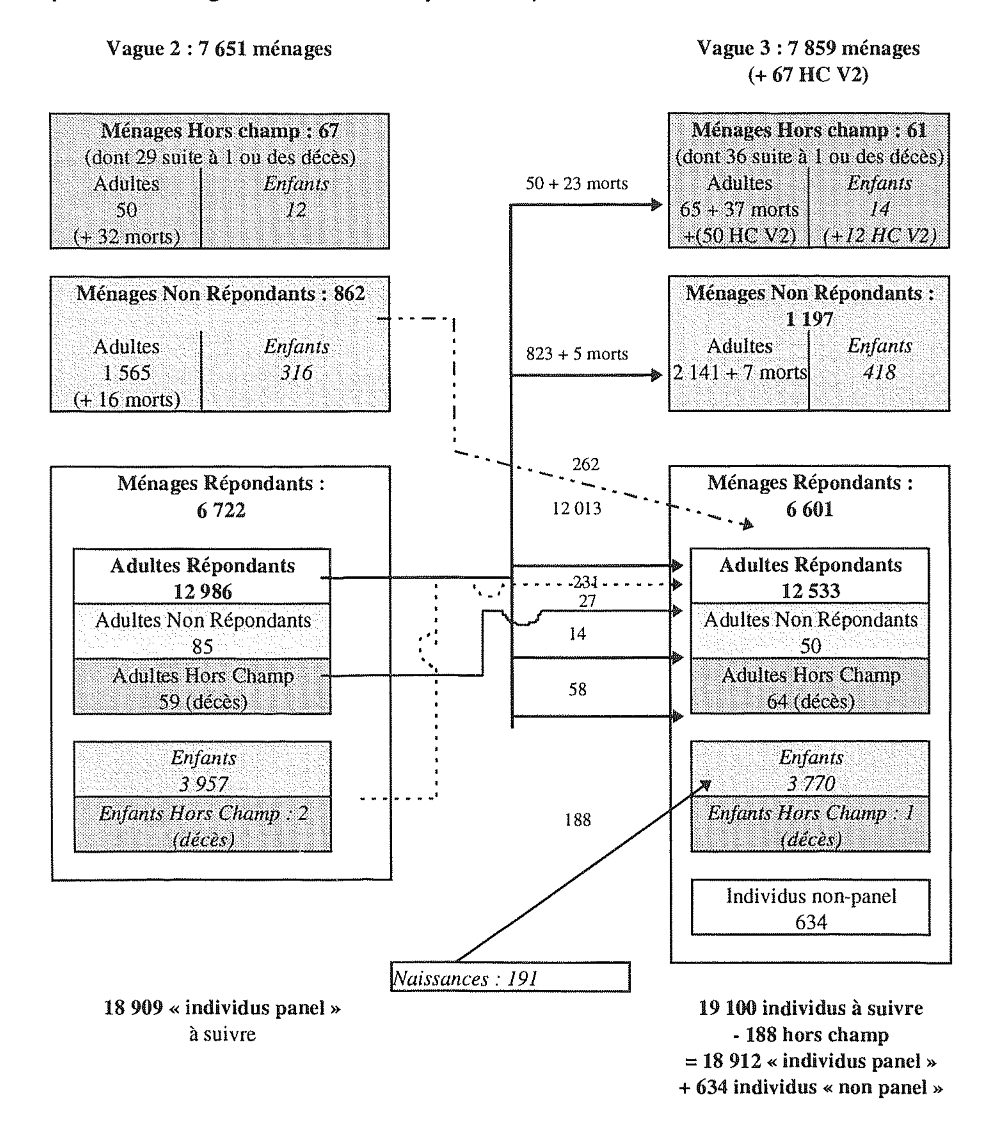


En vague 3 (V3), les 18 909 individus entrant dans le champ de V2 doivent théoriquement être suivis, qu'ils aient ou non répondu en V2. Des mouvements de sortie du champ de l'enquête se sont cependant produits entre V2 et V3 : 109 individus « panel » sont morts, 79 autres ont déménagé en ménage collectif ou à l'étranger. *A contrario*, 191 nouveau-nés de mère panel sont apparus. Au total, ce sont donc 18 912

individus panels à suivre, dont 14 724 adultes de 17 ans ou plus pour une interrogation individuelle. Parmi ces derniers, 260 avaient moins de 17 ans en V2.

Entre V2 et V3, 304 nouveaux ménages sont apparus par éclatement d'un ménage de V2. 67 ménages étant sortis du champ dès la vague 2, le nombre de ménages à contacter pour réaliser les entretiens individuels est donc désormais de 7 651-67+275=7 888, soit 3,1 % de plus qu'en V2 [voir schéma 2].

Schéma 2 : Evolution de l'échantillon entre les vagues 2 et 3 - Schéma simplifié (devenir et origine des adultes répondants)



Les taux de réponse sont encore relativement bons, même s'ils sont plus faibles

qu'en V2[5] : 6 601 ménages ont répondu à l'enquête, soit 84,6 % des 7 798 ménages entrant dans le champ de l'enquête. Au niveau individuel, 12 533 adultes panel (85,1 %) ont accepté l'interview. Ils représentent 96,5% des individus ayant répondu en vague 2. Parmi les ménages répondants, peu d'individus panel ont refusé de répondre (0,4 %) ; et la plupart avaient déjà refusé en V1.

Le taux d'attrition, défini comme l'écart à 1 du rapport entre le nombre d'individus panel répondants en V3 et le nombre d'individus panel répondants en V2 est de 3,5 %.

1.2. Choix d'un modèle de non-réponse individuelle

Avant de présenter le modèle retenu pour la correction de la non-réponse individuelle, il importe de bien préciser ce qu'on entend par non-réponse :

- la non-réponse ne concerne ici, rappelons-le, que les individus panel. Les individus non-panel ont par définition un poids de base longitudinal nul, le corriger n'aurait donc aucun sens...

- s'agissant des enfants, ils sont réputés avoir répondu dès lors que leur ménage a été contacté. Pour eux la correction de la non-réponse revient à corriger leur poids de base du taux de non-contact des ménages comprenant des enfants. Ce taux est de 1,29 % en V2, et de 1,75 % en V3 parmi les répondants de V2. Par convention, les nouveau-nés se voient affecter la moitié du poids de leur mère.

- au niveau des adultes (individus de 17 ans et plus), c'est l'existence d'un questionnaire individuel qui définit la réponse. Dans la suite, nous ne parlerons plus que de cette population.

Nous aurions pu estimer d'abord une probabilité de contact de l'individu à travers son ménage, puis une probabilité de non-réponse conditionnelle au contact. Nous avons préféré estimer directement la probabilité de non-réponse en introduisant dans le modèle explicatif des variables elles-mêmes fortement corrélées à la probabilité de contact : nationalité, type de ménage, catégorie socioprofessionnelle, voire déménagement du ménage, etc.

Deux modèles explicatifs de la non-réponse des individus panels adultes ont été estimés, en V3 comme en V2. Pour chacun d'entre eux, nous avons d'abord retenu comme variables exogènes des variables collectées lors de la vague d'enquête précédente, et donc disponibles à la fois pour les individus répondants et les individus non-répondants : sexe, âge en tranches, niveau d'études, type de ménage,

5. Les règles de suivi des individus, qui imposent de suivre les personnes ayant refusé de répondre en V2, engendrent mécaniquement ce taux plus élevé.

charges liées au logement, nationalité, activité et catégorie socioprofessionnelle de la personne de référence du ménage, type de commune, pauvreté monétaire du ménage. La différence entre les deux modèles porte sur l'introduction ou non d'une variable supplémentaire décrivant la mobilité géographique *depuis* la vague précédente.

Dans un premier temps, nous avons donc banni toutes les variables de comportement, telles que le déménagement ou l'éclatement du ménage. Nous considérions en effet qu'il était abusif de supposer une homogénéité de comportement entre les individus qui déménagent.

Dans un second temps, nous sommes revenus sur cette considération, et avons réintroduit les variables de mobilité géographique. Leur omission revenait en effet à supposer que les individus mobiles ne se distinguent pas des autres en terme de structure ni de comportement, une fois prises en compte quelques variables socio-démographiques de base. Cela constitue une hypothèse tout aussi forte, en particulier dans le cadre du panel dont un des principaux objectifs est justement d'étudier les corrélations entre différentes formes de mobilité (professionnelle, démographique, en terme de revenus,...).

Pour les vagues 2 et 3, les variables les plus discriminantes du point de vue de la non-réponse sont sensiblement les mêmes, soit :

- lorsqu'on estime les modèles sans variables de comportement : le type de ménage, la nationalité de la personne de référence du ménage, sa catégorie socioprofessionnelle et le type de commune, suivi par l'âge de l'individu en vague 2 ; la catégorie socioprofessionnelle de la personne de référence du ménage, sa nationalité, l'âge de l'individu, le type de ménage puis l'activité de la personne de référence et le poids des charges liées au logement... en vague 3 [tableau 1].

- lorsqu'on estime les modèles avec variables de comportement : le déménagement suite à un éclatement de ménage, le type de ménage, la nationalité de la personne de référence du ménage, sa catégorie socioprofessionnelle puis le type de commune... en vague 2 ; le déménagement suite à un éclatement de ménage, la catégorie socioprofessionnelle de la personne de référence du ménage, sa nationalité, le type de ménage, puis l'activité de la personne de référence et le poids des charges liées au logement... en vague 3 [tableau 2].

La variable de déménagement suite à un éclatement est de loin la plus discriminante. Les enquêteurs avouent d'ailleurs rencontrer beaucoup de difficultés pour obtenir l'adresse d'un enfant qui quitte le ménage, et dont les parents jugent que ce n'est pas la peine « d'aller l'embêter ». De même, les individus ne se montrent pas forcément coopératifs quand on leur demande les coordonnées d'un ex-conjoint. Elle semble toutefois moins discriminante en vague 3 qu'en vague 2.

Tableau 1 : Un premier modèle de non-réponse des adultes individus panel, sans variables de comportement

	VAGUE 2		VAGUE 3	
Variable	**Paramètre**	**Ecart-type**	**Paramètre**	**Ecart-type**
Constante	-2 0890	0 0645	-2.4121	0.1121
Homme	Référence			
Femme	-0,1207	0,0531	ns	ns
17 à 24 ans	0,3138	0,0731	Référence	
25-34 ans			0,2119	0,0968
35-54 ans	Référence		Référence	0,1226
55-64 ans			-0,3396	
65 ans et plus			Référence	
Niveau d'études atteint :				
CAP-BEP		0,0761	-0,1982	0,1083
autre, < à l'enseignement supérieur	Référence		Référence	0,1185
Enseignement supérieur			-0,2719	
	-0,2356			
Personne seule, couple sans enfant,				
famille monoparentale	Référence	0,0855	Référence	0,1222
Couple avec 2 enfants		0,1050	-0,3785	0,1154
Couple avec 1 enfant	-0,2750	0,0738	-0,2533	0,1279
Couple avec ≥ 3 enfants	-0,7214		-0,2910	
Autre ménage	0,1892		Référence	
Charges de logt supportables	Référence		Référence	
Charges de logt trop élevées	0,2144	0,0654	0,4233	0,0839
PR française	Référence			
PR européenne	0,4090	0,1296	Référence	0,1617
PR autre nationalité	0,5569	0,1285	0,6964	
Agglomération parisienne	0,3163	0,0730	0,2152	0,0989
Autre lieu	Référence		Référence	
Ménage non pauvre	Référence			
Ménage pauvre	0,2085	0,0785	non disponible	
PR agriculteur exploitant	Référence		-0,5356	0,2313
PR artisan, commerçant	0,5076	0,0885	0,4751	0,1474
PR cadre		0,0607	0,3266	0,1346
PR prof. intermédiaire	Référence		Référence	0,1170
PR employé			0,5155	0,1083
PR ouvrier	0,2662		0,2390	
PR non active occupée	Référence		Référence	
PR active occupée	-0,2700	0,0574	-0,4156	0,0836
PR a un CDD	0,2949	0,1454	ns	ns
Propriétaire du logement	ns	ns	-0,3745	0,0783
Autre statut			Référence	
Nombre d'observations	14 634		13 095	
-2*LogL	10 016 451		6 062 729	

PR = personne de référence du ménage

Tableau 2 : Un deuxième modèle de non-réponse des adultes individus panel intégrant des variables de comportement (déménagement, éclatement du ménage)

	VAGUE 2		VAGUE 3	
Variable	**Paramètre**	**Ecart-type**	**Paramètre**	**Ecart-type**
Constante	-2,1340	0,0561	-2,3992	0,1111
Niveau d'études atteint inférieur à l'enseignement supérieur	Référence	0,0779	Référence	0,1158
Enseignement supérieur	-0,3142		-0,2382	
Personne seule	Référence		Référence	
Couple sans enfant	Référence	0,0863	Référence	0,1250
Couple avec 1 enfant	-0,3224	0,1054	-0,4375	0,1191
Couple avec 2 enfants	Référence	0,1374	-0,3333	0,1316
Couple avec ≥ 3 enfants	-0,7498		-0,3886	0,1151
Famille monoparentale	-0,2914		Référence	
Autre ménage	Référence		-0,2976	
Charges de logt supportables	Référence		Référence	
Charges de logt trop élevées	0,2113	0,0670	0,4230	0,0845
PR française	Référence		)	
PR européenne	0,4198	0,1328	) Référence	0,1636
PR autre nationalité	0,5291	0,1333	0,7313	
Agglomération parisienne	0,3215	0,0746	0,2356	0,0994
Autre lieu	Référence		Référence	
Ménage non pauvre	Référence			
Ménage pauvre	0,2130	0,0804	non disponible	
PR agriculteur exploitant	Référence		-0,4987	0,2317
PR artisan, commerçant	0,5021	0,0901	0,4797	0,1480
PR cadre	)	0,0619	0,3476	0,1352
PR prof. intermédiaire	| Référence		Référence	0,1175
PR employé	)		0,5232	0,1085
PR ouvrier	0,2423		0,2638	
PR non active occupée	Référence		Référence	
PR active occupée	-0,2845	0,0583	-0,4326	0,0825
PR a un CDD	0,2764	0,1498	ns	ns
Propriétaire du logement	ns	ns	-0,3952	0,0774
Autre statut			Référence	
55-64 ans	ns	ns	-0,3263	0,1221
Autre tranche d'âge			Référence	
Déménagement suite à un éclatement de ménage	2,1048	0,1049	1,5826	0,1542
Autre cas	Référence		Référence	
Nombre d'observations	14 634		13 095	
-2*LogL	9 692 794			

PR = personne de référence du ménage

2. Poids de base des individus panel pour la vague 2

2.1 Correction de la non-réponse par groupes homogènes et calcul des poids de base pour la vague 2

La correction de la non-réponse a été effectuée en modifiant les poids de base V1 selon la formule :

$$\text{Poids de base V2} = \frac{\text{Poids de base V1}}{1 - \text{taux de non réponse}}$$

Afin de limiter la dispersion des poids résultant de la CNRGH, les taux de non-réponse ont été calculés pour des *catégories* homogènes définies par le croisement des modalités des variables apparues les plus discriminantes à l'étape précédente.

Les catégories ont ainsi été construites en croisant :

- dans le premier cas (pas de variables de comportement), le type de ménage, la nationalité de la personne de référence, sa catégorie socioprofessionnelle, le type de commune et l'âge de l'individu ;

- dans le second cas (intégration de variables de comportement) le déménagement suite à un éclatement de ménage, le type de ménage, la nationalité de la personne de référence du ménage, sa catégorie socioprofessionnelle et le type de commune.

Lorsqu'une catégorie comptait peu d'individus, on l'a regroupée avec la catégorie la plus proche pour le calcul du taux de non-réponse. La proximité entre catégories a été appréciée à partir des coefficients estimés dans le modèle. Au total, 31 catégories sont distinguées dans le premier cas, et 30 dans le second.

Les taux de non-réponse pour chacune de ces catégories sont présentés dans l'annexe 1, tableaux A et B. Ils varient, selon les catégories, entre 4,6 % et 38,8 % dans le premier cas (pas de variables de comportement), entre 2,9 % et 58,9 % dans le second (intégration de variables de comportement). La dispersion des poids devrait donc être plus importante dans le second cas.

2.2. Distribution des poids de base longitudinaux V2 des individus panel adultes

La distribution des poids de base des individus panel adultes en vague 2 est légèrement plus dispersée lorsqu'on intègre la variable de déménagement et d'éclatement de ménage à la grille de construction des catégories homogènes.

Un individu obtient un poids particulièrement élevé dans ce cas. Lorsqu'on supprime cet individu, la dispersion reste cependant légèrement plus forte que lorsque la CNRGH est réalisée sans variable de comportement. [tableau 3].

Nous reviendrons dans la partie 5 sur les caractéristiques de la population pour laquelle le rapport entre les deux séries de poids est très différent de 1.

Tableau 3 : Poids de base (corrigés de la non-réponse) des 12 986 individus panel répondants en Vague 2

	Présence de variables de déménagement et éclatement du ménage pour la CNRGH		Rapport entre ces deux séries de poids
	NON (1)	OUI (2)	(3)=(2)/(1)
Somme	45 178 331	45 178 331	
Moyenne	3 479,00	3 479,00	1,00070
Ecart-type	1 120,06	1 164,94	0,09631
Minimum	1 977,40	1 944,17	0,79964
1%	2 189,42	2 162,95	0,92087
5%	2 385,53	2 350,90	0,94971
10% D1	2 534,79	2 504,57	0,95951
25% Q1	2 817,04	2 803,93	0,97856
50% Med	3 264,46	3 248,95	0,98877
75% Q3	3 863,33	3 869,07	1,00898
90% D9	4 507,77	4 508,92	1,01317
95%	4 979,33	4 996,23	1,02216
99%	8 784,96	9 016,83	1,53573
Maximum	15 494,51	22 477,47 (*)	2,15742
Range	13 517,12	20 533,30	-
Q3-Q1	1 046,29	1 065,14	-
D9/D1	1,78	1,80	

(*) un seul individu a un poids si élevé. L'observation dont le poids est juste inférieur a un poids de 15395. (plus faible que le poids (1) maximum). Le rapport (2)/(1), pour cet individu au poids maximum, est de 1,498. La forte valeur de ce poids (2) maximum n'est donc que partiellement imputable à la grille de construction des catégories retenue. L'individu concerné (un homme de 30 ans, seul en vague 2 suite à un éclatement de son couple, vivant en agglomération parisienne) avait déjà un poids élevé en vague 1 : 13 297,86, la moyenne des poids étant alors de 3 109,97. Ce poids déjà fort en vague 1 tenait à deux facteurs : un poids de sondage déjà fort (logement ayant statut de résidence secondaire lors du Recensement de Population de 1990) combiné à une forte correction pour la non-réponse.

La comparaison des poids des vagues 1 et 2 montre également une amplitude des rapports de poids accrue lorsque la CNRGH intègre la variable de déménagement. Pour les individus panels adultes, ce rapport est en effet alors en moyenne de 1,1267, et varie entre 1,0297 et 2,4317[6], contre 1,1265 en moyenne lorsque la CNRGH n'intègre pas la variable de déménagement, le rapport étant alors compris entre 1,0486 et 1,6339.

3. Calage sur marges des poids de base de la vague 2

3.1. Population et variables de calage

Le calage sur marges a pour objet d'assurer une plus grande représentativité transversale de notre échantillon. Il a cependant été réalisé avant calcul des poids transversaux, conformément aux recommandations d'Eurostat.

Nous avons choisi de ne recaler que la population servant effectivement au calcul de ces poids transversaux : les adultes panel répondants. Nous avons opté pour un calage minimal, basé sur le respect des structures individuelles par âge décennal x sexe et du nombre d'individus par ménage, telles qu'estimées à partir de l'enquête Emploi de mars 1995.

Notre estimation donne une population plus importante que celle de l'enquête emploi. Cela peut s'expliquer par la convention adoptée pour la définition du champ de l'enquête : en l'absence d'éléments permettant de dire si un individu appartient au champ de l'enquête (parti sans laisser d'adresse), on considère qu'il en fait partie. Cela revient à surestimer le champ de l'enquête et le taux de non-réponse.

Avant calage, et quelle que soit la série de poids de base utilisée, l'échantillon du panel présente un déficit d'individus vivant dans des ménages d'1 ou 2 personnes, en particulier de femmes âgées de 72 ans ou plus.

6. la valeur maximale n'étant pas atteinte par l'individu au poids maximal.

3.2. Distribution des poids de base après calage

Tableau 4a : Distribution des poids des adultes panel répondants en vague 2 - Pas de variables de comportement (déménagement, éclatement) pour la CNRGH

	Poids avant CNRGH vague 2	Poids de base vague 2		Rapport des poids		Rapport des poids vague 2 avant et après calage
	(1)	non recalé (2)	recalé (3)	non recalé (4)=(2)/(1)	recalé (5)=(3)/(1)	(6)=(3)/(2)
Somme	40 048 604	45 178 331	44 842 091	-	-	-
Moyenne	3083,98	3479,00	3453,11	1,1265	1,1169	0,9912
Ecart-type	960,81	1120,06	1150,75	0,0601	0,0863	0,0501
Minimum	1867,58	1977,40	1868,69	1,0486	0,9258	0,8691
1%	1981,39	2189,42	2126,77	1,0486	0,9696	0,8957
5%	2152,39	2385,53	2318,54	1,0667	0,9941	0,9224
10%	2282,99	2534,79	2472,52	1,0741	1,0117	0,9273
25% Q1	2521,40	2817,04	2782,00	1,1007	1,0528	0,9506
50% Med	2904,91	3264,46	3215,17	1,1139	1,1148	0,9816
75% Q3	3421,74	3863,33	3836,20	1,1526	1,1725	1,0360
90%	3942,78	4507,77	4542,58	1,1613	1,2128	1,0667
95%	4327,9	4979,33	5049,71	1,2335	1,2497	1,0747
99%	7767,84	8784,96	8470,29	1,3211	1,3512	1,0916
Maximum	13885,4	15494,51	16376,98	1,6339	1,8431	1,0916
Range	12017,82	13517,12	14508,29	-	-	-
Q3-Q1	900,34	1046,29	1054,20	-	-	-
Mode	3022,08	4104,47	6478,70	1,1007	1,1403	1,0173

Le calage sur adultes répondants a été réalisé à l'aide du logiciel CALMAR. La méthode retenue est le Logit tronqué, les bornes retenues étant de 0,8 et 1,2, pour les deux calages réalisés.

Sans variables de comportement pour la CNRGH, les poids recalés se situent dans un rapport maximal de 1 à 8,8. Le rapport entre le premier et le 99ème centiles est de 4,0 ; entre le 5ème et le 95ème centiles, il tombe à 2,2. Le rapport entre les poids de base V2 avant et après calage en moyenne de 0,99 ; il varie entre 0,87 et 1,09 [tableau 4a].

Lorsqu'on intègre des variables de comportement pour la CNRGH (déménagement, éclatement de ménage), les poids recalés varient dans un rapport de 1 à 12,3. Le rapport entre le premier et le 99ème centiles est de 4,2 ; entre le 5ème et le 95ème centiles, il tombe à 2,2. Le rapport entre les poids de base V2 avant et après calage en moyenne de 0,99 ; il varie entre 0,90 et 1,10. [tableau 4b]. La distribution est donc ici légèrement plus dispersée (le rapport de 1 à 12,3 provenant, rappelons-le, du poids extrême d'un seul individu).

Tableau 4b : Distribution des poids des adultes panel - Intégration de variables de comportement pour la CNRGH

	Poids avant CNRGH vague 2	Poids de base vague 2		Rapport des poids		Rapport des poids vague 2 avant et après calage
	(1)	non recalé (2)	recalé (3)	non recalé (4)=(2)/(1)	recalé (5)=(3)/(1)	(6)=(3)/(2)
Somme	40 048	45 178 331	44 842 091	-	-	-
Moyenne	3083,98	3479,00	3453,11	1,1267	1,1176	0,9914
Ecart-type	960,81	1164,94	1185,72	0,1197	0,1335	0,0388
Minimum	1867,58	1944,17	1896,77	1,0297	0,9360	0,8979
1%	1981,39	2162,95	2116,00	1,0310	0,9703	0,9133
5%	2152,39	2350,90	2312,19	1,0511	1,0002	0,9377
10%	2282,99	2504,57	2454,37	1,0624	1,0138	0,9456
25% Q1	2521,40	2803,93	2771,16	1,0816	1,0559	0,9591
50% Med	2904,91	3248,95	3205,86	1,1105	1,1047	0,9889
75% Q3	3421,74	3869,07	3829,46	1,1334	1,1490	1,0142
90%	3942,78	4508,92	4517,62	1,1715	1,1915	1,0456
95%	4327,9	4996,23	5075,42	1,2084	1,2373	1,0724
99%	7767,84	9016,83	8877,38	1,6939	1,7171	1,0729
Maximum	13885,4	22477,47	23260,76	2,4317	2,5957	1,1004
Range	12017,82	20533,30	21363,99	-	-	-
Q3-Q1	900,34	1065,14	1058,30	-	-	-
Mode	3022,08	4195,42	6552,73	1,1105	1,0651	1,0361

4. Calcul des poids transversaux

4.1. La méthode employée : le partage des poids

Conformément aux recommandations d'Eurostat, c'est la méthode de partage des poids qui a été retenue pour le calcul des poids transversaux[7].

Ce principe se traduit par le partage entre tous les membres adultes du ménage de la somme des poids de base des adultes (personnes de 17 ans et plus), que ces derniers soient individus panel ou individus non-panel. Rappelons que les poids de base des individus non-panel sont nuls. Soit :

$$\text{poids transversal du ménage } i = w_i = \frac{1}{p+n}\sum_{k=1}^{p} u_{i,k}$$

d'adultes panel du ménage i et n le nombre d'adultes non-panel de ce où $u_{i,k}$ représente le poids de base en V2 de l'individu k dans le ménage i, p est le nombre ménage

Tous les membres adultes répondants d'un ménage reçoivent donc le même poids transversal, défini comme la moyenne de leurs poids de base.

4.2. Distributions au niveau ménage des poids transversaux V2[8] calculés sans ou avec calage préalable

Nous avons calculé des poids transversaux avant et après calage des poids de base V2 des adultes panel, pour chacune des deux séries de poids de base.

Parmi les 6.722 ménages répondants, 2 n'auront pas de poids transversal : l'un ne compte parmi ses trois membres qu'un seul individu panel, âgé de moins de 17 ans ; l'autre ne compte qu'un seul individu, qui n'a pas répondu au questionnaire individuel. Les distributions de poids sont donc données pour les 6 720 ménages pour lesquels leur calcul est possible [tableau 5].

Dans le cas d'un redressement sans intégration de variables de comportement lors de la CNRGH, on obtient les résultats suivants :

- avant calage, les poids transversaux se situent dans un rapport maximal de 1 à 24,8. Le rapport entre le premier et le 99ème centiles est de 6,7 ; entre le 5ème et le 95ème centiles, il tombe à 2,2.

7. pour une justification de la méthode, voir par exemple Lavallée (1995)
8. Le calcul des poids transversaux de la vague 3 repose sur le partage d'une série particulière de poids de base. Afin de ne pas perdre pour l'analyse les ménages dont aucun individu n'avait répondu en vague 2, ni n'avait donc de poids de base V3 (au sens employé jusqu'à présent), ces derniers ont été calculés par correction des poids de base V1. Nous ne les avons pour l'instant pas recalés, et nous n'en présenterons pas les distributions ici.

- après calage, le rapport maximal des poids transversaux est de 26,6, les rapports entre le premier et le 99ème centiles, et entre les 5ème et 95ème centiles restent sensiblement les mêmes (6,7 et 2,4).

Le rapport entre les poids transversaux avant et après calage est en moyenne de 1,00 ; il varie entre 0,88 et 1,09.

Dans l'autre cas, où la CNRGH intègre des variables de comportement, les poids transversaux sont davantage dispersés avant calage, mais le calage les modifie moins. Aussi, la dispersion après calage (rapport interdécile, notamment) est moins forte que dans le cas précédent :

- avant calage, les poids transversaux se situent dans un rapport maximal de 1 à 36,8. Le rapport entre le premier et le 99ème centiles est de 6,1 ; entre le 5ème et le 95ème centiles, il tombe à 2,3.

- après calage, le rapport maximal des poids transversaux est plus faible (35,7), les rapport entre le premier et le 99ème centiles, et entre les 5ème et 95ème centiles restent les mêmes (6,1 et 2,3).

Le rapport entre les poids transversaux avant et après calage est en moyenne de 0,998 ; il varie entre 0,91 et 1,07.

Tableau 5 : distribution au niveau ménage des poids transversaux de la vague 2

	Pas de variables de comportement (déménagement, éclatement) pour la CNRGH			Intégration de variables de comportement (déménagement, éclatement) pour la CNRGH		
	non calé	calé	calé/non calé	non calé	calé	calé/non calé
Somme	23 009 997	23 151 040	-	23 168 631	23 165 932	-
Moyenne	3 424,11	3 445,10	1,0040	3 447,71	3 447,31	0,9981
Ecart-type	1 202,007	1 253,73	0,0489	1 242,50	1 273,98	0,0354
Minimum	623,71	614,89	0,8847	610,14	651,27	0,9112
1%	1 399,69	1 411,89	0,9180	1 507,16	1 517,80	0,9334
5 %	2 241,24	2 178,26	0,9312	2 246,93	2 221,11	0,9449
10 % D1	2 441,11	2 395,54	0,9410	2 433,81	2 402,64	0,9523
25 % Q1	2 780,53	2 766,02	0,9640	2 777,34	2 753,24	0,9623
50 % Med	3 214,74	3 203,71	0,9951	3 218,69	3 193,80	0,9985
75 % Q3	3 827,15	3 861,97	1,0433	3 840,87	3 841,87	1,0238
90 % D9	4 489,01	4 588,56	1,0747	4 517,65	4 564,27	1,0443
95 %	5 012,58	5 155,74	1,0843	5 058,19	5 155,51	1,0729
99 %	9 412,26	9 390,29	1,0916	9 388,81	9 331,05	1,0729
Maximum	15 494,51	16 376,98	1,0916	22 477,47	23 260,76	1,0729
Range	14 870,80	15 762,09	-	21 867,33	22 609350	-
Q3-Q1	1 046,61	1 095,95	-	1 063,53	1 088,62	-
D9/D1	1,84	1,92	-	1,86	1,90	-

5. Incidence de la série de poids utilisée sur la valeur de quelques indicateurs

Les analyses présentées dans cette partie reposent essentiellement sur l'utilisation des différents jeux de pondérations calculés pour la vague 2 du panel. Les séries de poids calculées pour la vague 3 ne seront que peu mobilisées ici.

Cette partie comporte d'abord une rapide description des populations dont les poids sont le plus affectés par l'introduction ou non de la variable de déménagement lors de la correction de la non-réponse. Sont ensuite comparées diverses estimations : de la taille de la population, de sa structure et, données plus au coeur de l'enquête, des niveaux de vie et de la mobilité en terme d'activité. Ces dernières estimations sont déclinées sur la population entière et sur quelques segments de population particulièrement mobiles (jeunes, personnes seules...)

5.1 Population dont le poids de base est fortement affecté par l'introduction ou non des variables de déménagement et d'éclatement du ménage

Les poids de base de l'individu sont ici considérés comme fortement affectés par l'introduction des variables de comportement (déménagement, éclatement du ménage), s'ils diffèrent de plus de 5 %. 1,8 % des individus panels adultes ont un poids de base supérieur d'au moins 5 % lorsqu'on intègre les variables de déménagement. *A contrario*, 5,7 % ont un poids au moins 5 % plus faible.

Les individus dont le poids augmente fortement sont, en octobre 1995, dans près de 8 cas sur 10 des locataires (1/3 dans la population totale des individus panel adultes). 63 % sont des actifs occupés, 12,2 % des chômeurs, et 14,8 % des étudiants. 42 % d'entre eux ont changé de situation professionnelle entre octobre 1994 et octobre 1995. Parmi les actifs occupés, 1/3 ont un emploi à durée déterminée. Ils sont en effet plus jeunes que le reste de la population : 54 % ont moins de 25 ans, et 32 % entre 25 et 35 ans. Ce sont essentiellement des personnes seules (41,3 %) et des couples sans enfant (37,1 %). 1/3 a un niveau de diplôme supérieur au baccalauréat. Ils sont sur-représentés en milieu urbain, hors agglomération parisienne [tableau 6].

La population dont le poids est plus faible lorsqu'on intègre les variables de comportement est à la fois semblable et différente : moins souvent active occupée (43,4 %), mais plus souvent au chômage (8,5 %) que notre échantillon dans son ensemble. 17 % ont changé de situation professionnelle entre les deux vagues d'enquête, moins que le groupe précédent, mais plus que notre échantillon dans son ensemble. 25 % des actifs occupés occupent un CDD. Cette population est plus âgée

que la précédente, mais plus jeune que notre échantillon dans son ensemble : 42,8 % ont moins de 25 ans, et 15,2% entre 25 et 35 ans. Ce sont essentiellement des familles monoparentales (19,2%) et des ménages complexes (33,8%). 25% sont en formation ou en cours d'études initiales. Près d'un sur trois se trouve en agglomération parisienne.

Tableau 6 : Population dont le poids varie fortement selon le modèle de non-réponse

	Individus panel adultes dont le rapport poids avec variable de comportement / poids sans variable de comportement est ...		Ensemble de l'échantillon des individus panel adultes
	< 0,95	> 1,05	
Effectif	746	236	12 986
Part dans l'échantillon	5,7	1,8	100,0 %
Locataire	43,2	78,8	32,6
Propriétaire ou accéd.	51,5	12,7	61,9
Autre cas	5,3	8,5	5,5
Actif occupé	43,4	62,9	51,0
Chômeur	8,5	12,2	6,8
Etudiant	25,2	14,8	8,6
Retraité	12,6	3,0	21,6
Autre inactif	10,3	7,1	11,9
Changement de sit. professionnelle entre les deux enquêtes	17,8	42,2	11,9
Parmi les actifs, CDD	25,4	32,0	13,4
Moins de 25 ans	42,8	54,0	15,1
25-34 ans	15,2	32,1	18,2
35-44 ans	10,6	5,1	19,4
45-54 ans	10,0	5,0	16,7
55 ans et plus	21,4	3,8	30,6
Personne seule	7,9	41,3	13,0
Couple sans enfant	13,1	37,1	25,3
Couple avec enfant(s)	26,0	8,9	44,3
Famille monoparentale	19,2	7,2	4,5
Autres ménages	33,8	5,5	12,9
Commune rurale	21,6	13,5	28,7
U.U. < 20.000 hab.	12,6	20,3	12,2
UU 20.000-100.000	10,0	19,4	13,2
UU > 100.000 hab	24,3	33,3	27,2
agglomération Paris.	31,5	13,5	13,7

5.2 Taille de la population

Le panel estime une population d'individus trop nombreuse en vague 2, mais pas assez en vague 3, et une population de ménages sensiblement identique à celle de l'enquête Emploi de mars 1995 [tableau 7]. La prise en compte de l'éclatement des ménages lors de la CNRGH conduit à une estimation du nombre de ménages plus importante. Elle revient en effet à accorder un poids plus important aux individus ayant fondé un nouveau ménage, et gonfle donc mécaniquement le nombre de ménages. Les fusions de ménages panel étant extrêmement rares, rien ne vient compenser ce mouvement de création de nouveaux ménages.

Tableau 7 : Estimation de la taille de la population

	Nombre d'individus	Nombre de ménages
Vague 1	57 913 481	22 839 615
Vague 2 (poids transversal)		
1) pas de variables de comportement pour la CNRGH		
sans calage	57 553 918	23 009 997
avec calage	56 920 563	23 151 040
2) intégration de variables de comportement pour la CNRGH		
sans calage	57 464 226	23 168 631
avec calage	56 969 036	23 165 932
Vague 3 (poids transversal)		
1) pas de variables de comportement pour la CNRGH		
sans calage	55 853 357	23 220 977
2) intégration de variables de comportement pour la CNRGH		
sans calage	55 639 616	23 352 885
Enquête Emploi mars 1995	57 000 429	23 047 168

5.3 Eléments de structure de la population

Quel que soit le jeu de pondérations transversales adopté, le panel donne un peu moins d'étudiants, un peu plus d'enfants de moins de 15 ans. Peut-être y a-t-il simplement eu des erreurs sur l'enchaînement des questions lors du remplissage des questionnaires. Les jeunes semblent légèrement sous-représentés en structure dans le panel. On n'observe pas de différence majeure selon le système de poids utilisé.

En revanche, sans calage des poids, la proportion de ménages d'une personne est davantage sous-estimée lorsqu'on utilise la série de poids *sans* variable de comportement (27,6 %) que celle *avec* (28,2 %, contre 28,9 % dans l'enquête Emploi).

Le taux de propriétaires est légèrement surestimé dans le panel, par rapport à l'enquête Emploi, et d'autant plus qu'on utilise la série de poids sans variables de comportement. Les ménages de propriétaires sont en effet moins mobiles que les autres, donc plus souvent retrouvés lors de l'enquête. Leur poids relatif dans l'échantillon est d'autant plus grand qu'on ne prend pas en compte les déménagements [tableau 8].

Tableau 8 : Estimation de la proportion de ménages propriétaires de leur logement en vague 2

Système de pondérations transversales		Proportion de ménages propriétaires (en %)
Enquête emploi mars 1995		53,5
Panel vague 1		55,0
Prise en compte de variables de comportement pour la CNRGH et		
	- pas de calage	55,6
	- calage	55,3
Pas de variables de comportement pour la CNRGH et		
	- pas de calage	56,2
	- calage	55,6

5.4 Niveaux de vie et pauvreté

Lorsqu'on s'intéresse à la population dans son ensemble, les indicateurs usuels de distribution des niveaux de vie et de pauvreté apparaissent relativement peu affectés par le système de pondération utilisé [tableau 9]. Les inégalités et la pauvreté semblent légèrement plus importantes lorsqu'on intègre la variable de déménagement lors de la correction de la non-réponse, et que l'on recale l'échantillon, mais les différences restent relativement faibles (0,4 % sur le niveau de vie moyen, 0,3 points sur le taux de pauvreté). Cette constatation est plutôt rassurante.

Lorsqu'on décline les indicateurs selon des tranches d'âge, ou qu'on distingue le type de ménage, les différences sont en revanche plus marquées. Selon le système de poids retenu, l'estimation du taux de pauvreté des 15-29 ans varie ainsi entre 12,7 % et 13,3 %. De la même façon, l'estimation du taux de pauvreté des personnes seules est fortement liée au jeu de pondérations utilisé. L'attrition différentielle selon les catégories peut donc avoir des répercussions non négligeables sur les analyses de disparités intercatégorielles.

Tableau 9 : Quelques indicateurs transversaux de niveau de vie* et pauvreté des individus en 1994

	Pas de variables de comportement (déménagement, éclatement) pour la CNRGH		Intégration de variables de comportement (déménagement, éclatement) pour la CNRGH	
	sans calage	**avec calage**	**sans calage**	**avec calage**
Ensemble de la population				
- moyenne	103 717	103 603	103 421	103 282
- 1er décile (D1)	45 883	45 600	45 530	45 411
- 1er quartile (Q1)	62 430	62 320	62 320	62 155
- médiane (Med)	88 000	87 949	87 881	87 759
- 3ème quartile (Q3)	124 415	124 414	124 171	124 098
- 9ème décile (D9)	175 106	175 006	174 540	174 425
- rapport interdécile (D9/D1)	3,82	3,84	3,83	3,84
- indice de Theil	0,17117	0,17151	0,17156	0,17171
- variance des logarithmes	0,35268	0,35520	0,36322	0,36421
Taux de pauvreté (RUC<Med/2)	8,8	9,0	9,0	9,1
Individus				
- niveau de vie moyen	92 558	92 251	91 670	91 426
Taux de pauvreté (RUC<Med/2)	12,7	12,9	13,1	13,3
Personnes seules				
- niveau de vie moyen	91 801	91 733	90 771	90 669
Taux de pauvreté (RUC<Med/2)	17,1	17,2	17,8	17,8

*Le niveau de vie est défini comme le rapport entre le revenu total du ménage et son nombre d'unités de consommation (UC). L'échelle utilisée pour le calcul des UC accorde un poids de 1 au premier adulte du ménage, de 0,5 aux suivants et de 0,3 aux enfants de moins de 14 ans.
Un individu est dit « pauvre » si son niveau de vie est inférieur à la moitié du niveau médian. L'estimation du seuil de pauvreté varie donc avec l'estimation de la médiane, et se fait donc conditionnellement au système de pondération.

5.5 Trajectoire d'activité des individus panels adultes

Le panel européen permet de suivre mois par mois la trajectoire d'activité des individus adultes, à travers un calendrier d'activité. Ce calendrier permet de répertorier une vingtaine de types d'occupations. Concernant l'emploi, l'enquêté indique son statut (indépendant, salarié), le type de contrat qui le lie à son employeur (CDD, CDI) et son temps de travail (temps plein, temps partiel). Le calendrier ne permet donc pas de repérer un changement d'employeur, mais il indique les modifications survenues dans le statut de l'individu. A côté de l'emploi sont distinguées les périodes de chômage, de maladie, d'études ou de formation, de retraite, etc. L'exploitation de ce calendrier a conduit à définir 19 types de trajectoires sur 12 mois, puis sur 24 mois, selon le nombre de changements intervenus au cours de la période et le type d'activité principale entre deux changements. Nous ne présenterons ici que les trajectoires sur 12 mois les plus fréquentes, ou celles dont l'estimation de la fréquence est fortement influencée par le système de pondérations longitudinales retenu [tableau 10a].

Globalement, le choix du système de poids longitudinaux ne semble pas influer sur la répartition estimée de la population entre les différentes trajectoires d'activité définies. Quel que soit le système de poids, on constate une surestimation, par rapport à la vague 1 du panel, de la proportion de personnes constamment en formation en 1993 ; la proportion estimée de retraités est au contraire plus faible. Pour les autres trajectoires d'activité, les proportions d'individus concernés coïncident très exactement d'une estimation à l'autre, quelles que soient la série de poids et la vague d'enquête utilisées.

Lorsqu'on examine ces trajectoires d'activité selon l'âge de l'individu, le diagnostic reste identique [tableau 10b].

Lorsqu'on construit les trajectoires à partir des observations relatives aux 24 mois de 1993 et 1994, on retrouve les résultats présentés ici sur des trajectoires construites à partir des seules observations de l'année 1993.

Tableau 10a : Trajectoires d'activité des individus panel au cours de l'année 1993

	Population et poids de la vague 1	Pas de variables de comportement (déménagement, éclatement) pour la CNRGH		Intégration de variables de comportement (déménagement, éclatement) pour la CNRGH	
		sans calage	avec calage	sans calage	avec calage
Aucun changement en 1993					
- CDI temps plein	32,5	32,7	32,7	32,6	32,6
- CDI temps partiel	3,0	3,1	3,0	3,1	3,0
- CDD temps plein	1,4	1,4	1,4	1,4	1,4
- indépendant	5,3	5,1	5,1	5,1	5,1
- chômage	2,9	2,6	2,7	2,7	2,7
- retraite	23,5	22,7	23,0	22,6	23,0
- formation	19,6	20,7	20,4	20,8	20,5
Au moins 1 changement en 93					
- passage au chômage	2,1	2,1	2,1	2,1	2,1
- sortie de chômage	1,4	1,4	1,4	1,4	1,4
- autres trajectoires	7,1	6,8	6,8	6,8	6,8
Ensemble	*100,0*	*100,0*	*100,0*	*100,0*	*100,0*

Les trajectoires présentées ici font état des changements de statut des personnes sur le marché du travail. Les changements de profession ou d'employeur leur échappent. Or, on peut penser qu'un certain nombre de déménagements sont liés à de tels changements. Une question du panel permet de repérer les individus dont la situation (profession, employeur) à la date d'enquête n'est plus la même que celle relevée lors de l'enquête précédente. On note alors que l'estimation de la proportion d'individus concernés est un peu plus sensible au choix du système de poids : sans intégration de la variable de déménagement lors de la correction de la non-réponse, cette proportion est évaluée à 11,9 % ; lorsqu'on intègre cette variable, l'estimation passe à 12,2 %. Appliqué aux données de la vague 3 du panel, l'exercice conduit à

peu de choses près au même résultat, avec des proportions estimées respectivement à 11,6 % et 11,8 %. Les différences constatées en vague 2 ne sont pas amplifiées en vague 3, où l'attrition, il est vrai, est de plus faible ampleur.

Tableau 10b : Trajectoires d'activité au cours de l'année 1993 des individus panel de moins de 30 ans

	Population et poids de la vague 1	Pas de variables de comportement (déménagement, éclatement) pour la CNRGH		Intégration de variables de comportement (déménagement, éclatement) pour la CNRGH	
		sans calage	avec calage	sans calage	avec calage
Aucun changement en 1993					
- CDI temps plein	23,4	22,7	23,3	22,9	23,2
- CDI temps partiel	1,8	1,8	1,8	1,8	1,8
- CDD temps plein	3,0	2,9	2,9	2,9	2,9
- indépendant	1,1	1,3	1,3	1,3	1,3
- chômage	2,8	2,3	2,3	2,4	2,8
- formation	45,9	48,7	47,9	48,5	48,0
Au moins 1 changement en 93					
- passage au chômage	3,8	3,6	3,7	3,6	3,6
- sortie de chômage	2,4	2,3	2,4	2,4	2,4
- autres trajectoires	15,8	14,4	14,4	14,2	14,0
Ensemble	*100,0*	*100,0*	*100,0*	*100,0*	*100,0*

Conclusion

Le calcul des pondérations dans le cadre d'un panel implique de suivre les individus de façon précise et de bien identifier les caractéristiques des non-répondants.

Le choix du modèle de non-réponse peut influer fortement sur les distributions de poids. Cependant, lorsqu'on compare les résultats d'analyses effectuées avec deux séries de pondérations issues de modèles différents, il s'avère que les résultats sont globalement proches. Les poids construits dans chacun des systèmes ne divergent en effet fortement que pour des segments très particuliers, et peu nombreux, de la population.

Les statistiques produites sur ces sous-populations particulières semblent, en revanche, être affectées par ce choix. Les écarts constatés sont statistiquement non significatifs mais, en l'absence de calcul de précision des estimations effectuées, ils peuvent conduire à des interprétations contradictoires. Supposons par exemple que le taux de pauvreté en 1993 était de 8,7 %. Selon le système de poids retenu, on conclura tantôt à une stabilité (8,8 %), tantôt à un accroissement (9,1 %) de la proportion de pauvres...

Eléments de bibliographie

CASES C., « Méthodologie du panel européen de ménages : exploitation des données de la vague 1 du fichier français », *Document de Travail de la Direction des Statistiques Démographiques et Sociales*, n° F9705, Insee, 1997.

CHAMBAZ C., SAUNIER J.-M., VALDELIEVRE H., « Méthodologie du panel européen de ménages : exploitation des données de la vague 2 du fichier français », *Document de Travail de la Direction des Statistiques Démographiques et Sociales*, n° F9715, Insee, décembre 1997.

DEVILLE J.-C., « Les enquêtes par panel : en quoi diffèrent-elles des autres enquêtes ? suivi de Comment attraper une population en se servant d'une autre ? », *Journées de Méthodologie Statistique,* 1998, à paraître.

EUROSTAT, « Groupe de Travail « Panel Communautaire de Ménages », Paris 18 et 19 Septembre 1995, Pondération Longitudinale », *Doc. PAN 51/95,* Eurostat, juillet 1995.

LAVALLEE P., « Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode de partage des poids », *Techniques d'enquête*, Vol.21, n°1, Statistique Canada, juin 1995.

LEGENDRE N., « Calcul des pondérations du fichier français de la vague 3 du panel européen de ménages », *Note interne*, 1998, à paraître.

Correction de la non-réponse par catégories homogènes : taux de non-réponse par catégories

Tableau A : Taux de non-réponse par catégories retenus pour le calcul des poids de base Vague 2 - Pas de variables de comportement (déménagement, éclatement du ménage)

					Effectif	Taux de non-réponse
Enfants individus panel					4 062	0,0129
Adultes individus panel					14 632	0,1144
Couples + 1 enfant	PR française	PR artisan	-	-	180	0,1740
		PR ouvrier	-	< de 25 ans	111	0,1023
				≥ 25 ans	536	0,0797
		PR autre CS	Hors aggl. Paris	< 25 ans	161	0,0828
				≥ 25 ans	823	0,0690
			Agglo. Paris	-	186	0,0563
	PR	-	-		104	0,1686
Couples + ≥ 3 enfants	PR française	PR artisan	-	-	89	0,0625
		PR ouvrier	-	< 25 ans	190	0,0929
				≥ 25 ans	430	0,0679
		PR autre CS	Hors aggl. Paris	< 25 ans	219	0,0693
				≥ 25 ans	476	0,0464
			Agglo. Paris	-	105	0,0676
	PR	-	-	-	262	0,1164
Ménages complexes	PR française	PR artisan	-	-	171	0,1893
		PR ouvrier	-	-	566	0,1128
		PR autre CS	-	< 25 ans	126	0,1606
				≥ 25 ans	801	0,1191
	PR	PR art. / ouvri.	-	-	117	0,3880
		PR autre CS	-	-	60	0,1233
Autres ménages (*)	PR française	PR artisan	Hors aggl. Paris	-	649	0,1389
			Agglo. Paris	-	94	0,2430
		PR ouvrier	-	< 25 ans	290	0,1611
			Hors aggl. Paris	≥ 25 ans	2 105	0,1324
			Agglo. Paris	-	132	0,1994
		PR autre CS	Hors aggl. Paris	< 25 ans	562	0,1239
				≥ 25 ans	3 802	0,0914
			Agglo. Paris	< 25 ans	124	0,1279
				≥ 25 ans	767	0,1136
	PR	PR art. / ouvri.	-	-	282	0,2132
		PR autre CS	-	-	112	0,1912

PR = personne de référence du ménage
(*) : personnes seules, couples sans enfants, couples avec 2 enfants, familles monoparentales

Tableau B : Taux de non-réponse par catégories retenus pour le calcul des poids de base Vague 2 - intégration de variables de comportement (déménagement, éclatement du ménage)

						Effectif	Taux de non-réponse
Enfants individus panel						4 062	0,0129
Adultes individus panel						14 632	0,1144
Pas de déménagement ou alors non consécutif à un éclatement de ménage	Couple + 1 enfant	PR française	PR artisan	-	-	176	0,1724
			PR ouvrier	-	-	630	0,0692
			PR autre CS	Hors agg Paris	≤ bac.	669	0,0751
				.	> bac.	288	0,0486
				Agglo Paris	≤ bac.	90	0,0350
					> bac.	87	0,0288
		PR étrangère	-	-	-	100	0,1096
	Couple + ≥ 3 enfants	PR française	PR artisan	-	-	87	0,0343
			PR ouvrier	-	-	588	0,0562
			PR autre CS	Hors agg Paris	≤ bac.	452	0,0588
				.	> bac.	215	0,0301
				Agglo Paris	-	105	0,0676
		PR étrangère	-	-	-	255	0,0935
	Famille mono-parentale	-	PR artisan ou ouvrier	-	-	148	0,1513
		-	PR autre CS	Hors agg Paris	-	378	0,0754
				Agglo Paris	-	98	0,0829
	Autre type de ménage	PR française	PR artisan	Hors agg Paris	-	767	0,1464
				Agglo Paris	-	107	0,2239
			PR ouvrier	Hors agg Paris	≤ bac.	2 590	0,1177
					> bac.	115	0,1021
				Agglo Paris	-	186	0,1466
			PR autre CS	Hors agg Paris	≤ bac.	3 461	0,0995
				.	> bac.	1 172	0,0814
				Agglo Paris	≤ bac.	581	0,1329
					> bac.	360	0,0792
		PR étrangère	PR ouvrier ou artisan	-	-	353	0,2346
			PR autre CS	-	-	139	0,1217
déménagement suite à un éclatement de ménage	Couple + 1 ou ≥3 enfants et famille .monopar.	-	-	-	-	171	0,4096
	Autre type de ménage	-	PR artisan ou ouvrier	-		131	0,5888
		-	PR autre CS	-		134	0,4084

PR = personne de référence du ménage

Calcul des poids longitudinaux en vague 3

La correction de la non-réponse a été effectuée en modifiant les poids de base V2 selon la formule :

$$\text{Poids de base V3} = \frac{\text{Poids de base V2}}{1 - \text{taux de non réponse}}.$$

Les catégories homogènes par rapport à la non-réponse ont été construites en croisant :

- dans le cas où la CNRGH n'intègre pas de variables de comportement : la catégorie socioprofessionnelle de la personne de référence, sa nationalité, l'âge de l'individu et le poids des charges de logement lorsque les effectifs concernés le permettaient. Un nombre très important de catégories homogènes (47) a été défini, les taux de non-réponse variant entre 0,7% et 17,3%. Certaines, très proches, auraient pu être regroupées.

- dans le cas contraire (intégration de variables de comportement) : le déménagement suite à un éclatement de ménage, la catégorie socioprofessionnelle de la personne de référence du ménage, sa nationalité et le type de ménage. 33 catégories ont ainsi été définies, avec des taux de non-réponse plus différenciés, compris entre 1,9% et 31,5%.

Distribution des poids de base V3 des individus panel adultes

L'objectif étant de suivre des trajectoires sur 3 ans, les poids de base V3 ont été calculés pour les seuls individus panel ayant répondu à la fois en vague 1 et en vague 2. Ceux qui n'avaient pas répondu en vague 2 avaient en effet un poids de base nul à cette date. Leur poids de base V3 était donc également nul.

Ce choix correspond à une restriction du champ de l'enquête, et conduit donc à éliminer des analyses tous les ménages répondant en vague 3 mais dont aucun individu n'avait répondu en vague 2. Une deuxième série de poids de base V3 a donc été calculée par correction non plus des poids de base V2 mais des poids de base V1. Nous ne présentons pas le modèle de correction ici, ni la distribution des poids qui en découle.

Tableau C : Poids de base (corrigés de la non-réponse) des 12 244 individus panel répondants en vague 3 et ayant répondu en vague 2

	Présence de variables de déménagement et éclatement du ménage pour la CNRGH		Rapport entre ces deux séries de poids
	NON (1)	OUI (2)	(3)=(2)/(1)
Somme	45 354 696	45 314 628	
Moyenne	3 704,24	3 700,97	1,00073
Ecart-type	1 229,89	1 273,76	0,10499
Minimum	2 056,79	2 010,54	0,72250
1%	2 313,79	2 265,97	0,85901
5%	2 498,77	2 474,24	0,93200
10% D1	2 660,57	2 632,96	0,95035
25% Q1	2 983,94	2 968,24	0,97100
50% Med	3 454,61	3 438,65	0,98875
75% Q3	4 110,09	4 104,66	1,02723
90% D9	4 841,15	4 846,42	1,00865
95%	5 404,15	5 478,19	1,06207
99%	9 553,46	9 682,93	1,52679
Maximum	16 620,98	24 438,25	2,25961
Range	14 564,19	22 427,71	-
Q3-Q1	1 126,14	1 136,42	-
D9/D1	1,82	1,84	

En vague 3 comme en vague 2, la distribution des poids de base des individus panel adultes est légèrement plus dispersée lorsqu'on intègre la variable de déménagement et d'éclatement de ménage à la grille de construction des catégories homogènes.

L'individu dont le poids était particulièrement élevé dans ce cas en vague 2 conserve un poids beaucoup plus gros que l'individu qui le suit immédiatement, mais sans accroître la distance. Lorsqu'on supprime cet individu, la dispersion est toujours légèrement plus forte que lorsque la CNRGH est réalisée sans variable de comportement. [tableau C]

La comparaison des poids des vagues 2 et 3 montre là encore une plus grande déformation des poids lorsqu'on intègre les variables de déménagement et d'éclatement de ménage. Sans ces variables, le rapport des poids V3/V2 s'établit en moyenne à 1,0696, variant entre 1,0072 et 1,2090. Avec ces variables, le rapport moyen V3/V2 est légèrement plus faible (1,0693), mais sa distribution couvre un intervalle plus large ([1,0192 ;1,4590]). Le diagnostic est le même lorsqu'on compare les poids des vagues 1 et 3.

Session 2

Collecte et enquêteurs

UNE MÉTHODE DE MESURE DE L'EFFET ENQUÊTEUR

Catherine Berthier, Jean-Claude Deville, Bernard Néros

Le contexte des enquêtes auprès des ménages

La plupart des enquêtes auprès des ménages à l'Insee se déroulent en face à face par visite d'enquêteur. L'échantillon propre à une enquête est issu d'une base de logements, par un plan de sondage complexe, à plusieurs degrés, qui aboutit à confier à chaque enquêteur des logements pas trop éloignés les uns des autres, au nombre de 25 environ pour un échantillon de taille moyenne (soit une dizaine de milliers de logements).

Pour une enquête donnée, on cherche s'il existe une part de variation des réponses due au fait que tous les ménages ne sont pas interviewés par le même enquêteur. Autrement dit, l'effet enquêteur existe si un ménage est susceptible de donner des réponses différentes à des enquêteurs différents. A défaut d'interroger deux fois chaque ménage, on mesure cette variation sur deux échantillons de ménages équivalents, qui ont répondu à des enquêteurs différents pour un même questionnaire. La première vague du panel européen a fourni la possibilité de mettre en place cette expérience.

Il s'agit d'une mesure globale car elle ne conduit pas à isoler le travail d'un enquêteur particulier : on s'intéresse à la différence entre les résultats tirés de deux enquêtes en tout point semblables, excepté le fait qu'elles ont été menées par deux groupes d'enquêteurs.

Cette préoccupation autour de l'effet enquêteur redevient actuelle au moment où la collecte assistée par ordinateur (système Capi) se généralise. Cette transformation du mode de collecte peut conduire à faire intervenir un nombre plus limité d'enquêteurs, parce que leur formation doit être accrue. Or cette concentration joue sur l'effet enquêteur, et on court le risque qu'elle ne l'amplifie.

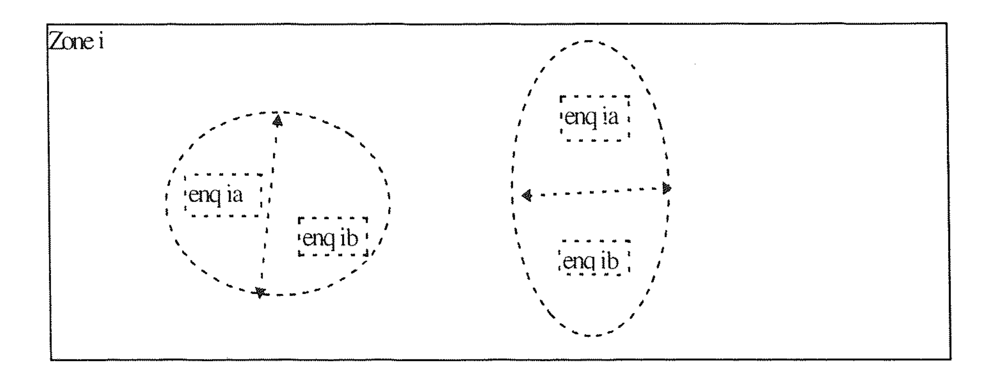
La première vague du panel européen ne s'est pas déroulée sous le système Capi, mais de manière traditionnelle, par remplissage de questionnaires sous forme papier. Un certain nombre d'erreurs peuvent subsister, commises par l'enquêteur, (filtres mal respectés, modalités impossibles....) qui seraient évitées dans une collecte sous Capi. L'effet enquêteur n'est pas entaché par ce type d'erreurs, puisqu'on le mesure à partir des données apurées.

Le dispositif expérimental

Un dispositif spécifique a été mis en place sur la première vague de l'enquête panel européen, permettant d'isoler l'effet enquêteur en neutralisant l'effet du terrain d'enquête. Un certain nombre de zones ont été sélectionnées avant le tirage de l'échantillon propre à l'enquête, chaque zone regroupant deux terrains couverts par un lot de deux enquêteurs. Dans chaque zone, les deux terrains ont été jugés ressemblants et présentant une homogénéité interne assez forte. Les zones sont pour la plupart urbaines parce qu'on y trouve plus facilement des terrains équivalents.

Les deux enquêteurs correspondant à une zone se sont partagé par moitié les échantillons des deux terrains. L'opération a porté sur 58 zones et 116 enquêteurs, pour 1400 ménages.

Graphique 1 :
Le partage d'une zone sur une paire d'enquêteurs désignés par enq ia et enq ib



Estimation de l'effet enquêteur

Grâce à ce dispositif qui respecte l'organisation de la collecte, on considère disposer de I zones sur lesquelles 2I échantillons de ménages indépendants ont été répartis aléatoirement sur 2I enquêteurs. On peut donc former deux groupes d'enquêteurs, notés groupe a et groupe b, qui ont mené deux enquêtes en parallèle dans l'ensemble des I zones.

Graphique 2 :
Deux enquêtes en parallèle

	groupe a		groupe b
zone i	*enquêteur 'ia'* ménage n°1 . . . .. ménage n°19 (n_{ia}=19)		*enquêteur 'ib'* ménage n°20 . .. . ménage n°35 (n_{ib}=16)

Dans chaque zone i, on dispose des deux enquêtes a et b. Pour une variable d'intérêt X, on estime la moyenne dans chaque zone $\overline{X}_{ia}$ et $\overline{X}_{ib}$, à l'aide de deux échantillons, de taille n_{ia} et n_{ib}. Au mieux, on dispose de la mesure de X sur tous les ménages des deux échantillons, qu'on a choisis de même taille. Mais le plus souvent il y a non-réponse totale et non-réponse partielle, et n_{ia} et n_{ib} sont les tailles en général différentes des deux sous-échantillons de répondants propres à X. Lorsqu'on estime l'effet enquêteur à partir des seuls répondants, on se limite au cas où la non-réponse obéit à un mécanisme homogène à l'intérieur de la zone.

En l'absence d'effet, on noterait m_i et σ_i^2 la moyenne et la variance de X dans la zone i. S'il y a effet enquêteur, cette moyenne et cette variance intègrent une erreur de mesure individuelle propre à chaque enquêteur. On modélise cette erreur de la manière suivante :

$$E\left(\overline{x}_{ia} / ia \text{ est tiré}\right) = m_i + \mu_{ia} \quad E\left(\overline{x}_{ib} / ib \text{ est tiré}\right) = m_i + \mu_{ib}$$

$$V\left(\overline{x}_{ia} / ia \text{ est tiré}\right) \approx \frac{\sigma_{ia}^2}{n_{ia}} \qquad V\left(\overline{x}_{ib} / ib \text{ est tiré}\right) \approx \frac{\sigma_{ib}^2}{n_{ib}}$$

(on assimile le sondage à un sondage aléatoire simple à taux de sondage négligeable)

Les μ_{ia} et μ_{ib} sont 2I variables aléatoires indépendantes de moyenne nulle, parce qu'on ne peut pas identifier de biais, et de variance commune σ^2_{enq}.

On cherche à estimer σ^2_{enq} :

En constatant que $E\left(\bar{x}_{ia} - \bar{x}_{ib} / _{ia,ib}\right) = \mu_{ia} - \mu_{ib}$, et en utilisant la formule approchée de variance ci-dessus, il vient :

$$(\mu_{ia} - \mu_{ib})^2 = E\left((\bar{x}_{ia} - \bar{x}_{ib})^2 / _{ia,ib}\right) - \frac{\sigma^2_{ia}}{n_{ia}} - \frac{\sigma^2_{ib}}{n_{ib}}$$

Or $(\mu_{ia} - \mu_{ib})$ est de moyenne nulle et de variance $2\sigma^2_{enq}$.

On en déduit que, en notant s^2_{ia} et s^2_{ib} les variances corrigées, dans chaque zone i, $(\bar{x}_{ia} - \bar{x}_{ib})^2 - \frac{s^2_{ia}}{n_{ia}} - \frac{s^2_{ib}}{n_{ib}}$ estime sans biais $2\sigma^2_{enq}$.

Un estimateur sans biais de σ^2_{enq} est donc :

$$\boxed{\hat{\sigma}^2_{enq} = \frac{1}{2I} \sum_{i=1}^{I} \left[(\bar{x}_{ia} - \bar{x}_{ib})^2 - \frac{s^2_{ia}}{n_{ia}} - \frac{s^2_{ib}}{n_{ib}}\right]}$$

Cet estimateur de variance peut conduire à des valeurs négatives.

Comment tester l'existence de l'effet enquêteur

L'absence d'effet enquêteur revient à : $\sigma^2_{enq} = 0$

On cherche à tester la nullité de σ^2_{enq}, pour accepter ou rejeter l'existence d'un effet enquêteur, contre l'hypothèse alternative $\sigma^2_{enq} > 0$. Mais on ne dispose pas de la loi de l'estimateur $\hat{\sigma}^2_{enq}$ sous l'hypothèse d'absence d'effet enquêteur.

Pour passer outre, on *simule* la loi de $\hat{\sigma}^2_{enq}$, sous l'hypothèse privilégiée, celle d'absence d'effet enquêteur. Sous cette hypothèse, les deux groupes a et b ont les

mêmes propriétés statistiques que d'autres groupes de ménages, pourvu que ces groupes présentent la même répartition par zones.

Pour obtenir ces autres groupes de ménages, l'idée est de réutiliser les mêmes données d'enquêtes, collectées sur les mêmes ménages, en *mélangeant* ces ménages pour reformer deux autres groupes. Ces mélanges s'effectuent en respectant les frontières des zones : à partir des $n_{ia} + n_{ib}$ ménages de la zone i, on peut former par mélange aléatoire deux nouveaux groupes i1 et i2, qui ne respectent plus la séparation par enquêteur, tout en obéissant à la contrainte de répartition, n_{ia} ménages d'un côté, et n_{ib} de l'autre.

On choisit par exemple de classer tous les ménages enquêtés de la zone i suivant un aléa, on affecte les données collectées auprès des n_{ia} premiers ménages au groupe i1, et les suivantes au groupe i2. On empile ensuite ces mélanges sur les I zones et on calcule à partir des deux nouveaux groupes 1 et 2 la statistique

$$\frac{1}{2I} \sum_{i=1}^{I} \left[\left(\overline{x}_{i1} - \overline{x}_{i2} \right)^2 - \frac{s_{i1}^2}{n_{i1}} - \frac{s_{i2}^2}{n_{i2}} \right]$$

On obtient *une autre* valeur de l'estimateur de variance.

Pour obtenir toute une distribution, on crée autant de fois qu'on veut deux groupes. Mille « mélanges » de ce type simulent une distribution de cet estimateur de variance.

On dispose de *la valeur observée* $\hat{\sigma}_{enq}^2$, celle calculée à partir des deux groupes d'enquêteurs a et b. On teste la nullité de σ_{enq}^2 en plaçant cette valeur observée dans la distribution simulée. Le fractile auquel correspond la valeur observée fournit une p-valeur à partir de laquelle on accepte ou on rejette l'hypothèse d'absence d'effet enquêteur.

Plutôt que de retenir pour une variable X donnée la seule conclusion du test pour un niveau fixé, on préfère conserver le fractile auquel correspond la valeur observée : si celui-ci est proche de la médiane, on conclut à l'absence d'effet enquêteur. Si ce fractile est décalé vers la droite de la distribution, correspondant à une valeur élevée, (il s'agit d'un test unilatéral), on soupçonnera un effet enquêteur. Si on constate un décalage pas très fort, mais se produisant sur plusieurs variables, on y verra une confirmation de l'effet enquêteur, alors qu'en fixant un niveau de test, on aurait rejeté l'hypothèse d'effet enquêteur sur chacune des variables examinées séparément.

En outre, ce procédé de test se justifie parce que *s'il y a effet enquêteur*, le fait de mélanger les ménages pour qu'ils ne soient plus regroupés par groupes d'enquêteurs, *atténue* l'effet enquêteur, mais ne l'annule pas complètement. De ce

fait en comparant la vraie valeur à la distribution simulée, on sous-estime le décalage dû à l'effet enquêteur.

Comment figurer une différence entre les statistiques provenant des deux groupes d'enquêteurs

Pour voir une différence éventuelle entre les statistiques tirées de chacune des enquêtes menées par les deux groupes d'enquêteurs, on étend cette méthode à une fonction T des observations. T peut être un total, une moyenne, un quantile, un ratio, une variance..... A partir de groupes i1 et i2 obtenus selon le même principe de mélanges, on calcule les statistiques T_{i1} et T_{i2}. Du point de vue de cette statistique, une distance $D_{1,2}$ entre les deux groupes 1 et 2 formés en empilant les groupes i1 et i2 peut être évaluée comme : $D_{1,2} = \sum_{i=1}^{I} |T_{i1} - T_{i2}|$

(On pourrait d'ailleurs généraliser en introduisant $\sum_{i=1}^{I} f(|T_{i1} - T_{i2}|)$, f possédant certaines propriétés)

La distribution de $D_{1,2}$ est simulée en formant mille fois deux groupes 1 et 2, par mélanges.

On examine comment se situe sur la distribution *la valeur observée,* celle qui mesure la distance entre les deux groupes d'enquêteurs $D_{a,b} = \sum_{i=1}^{I} |T_{ia} - T_{ib}|$

Le fait de former deux groupes par hasard, et non plus par enquêteur, entraîne une compensation des effets individuels des deux enquêteurs à l'intérieur de chaque groupe. On attribue donc à l'effet enquêteur le fait que la distance entre les deux groupes d'enquêteurs paraisse élevée dans la distribution des distances entre groupes formés par hasard.

Application de la méthode

Cette méthode a tout d'abord été appliquée sur la phase qui précède l'interview, qui comprend le repérage et le contact des deux enquêtes.

L'information tirée de cette phase a été résumée d'une part par la distinction entre répondants et non-répondants à l'enquête (pour les logements déclarés dans le champ de l'enquête), d'autre part par le statut de l'enquête (cette fois pour l'ensemble des logements de l'échantillon).

Extrait du questionnaire de l'enquête panel européen

Le statut de l'enquête

L'enquêteur doit placer le ménage dans l'une des catégories suivantes :

- le logement est hors champ
- le logement est dans le champ, mais on n'obtient pas de réponse du ménage.
- le logement est dans le champ, le ménage accepte de répondre à l'enquête.

Le montant du revenu :

1- Pour résumer :

En considérant l'ensemble des revenus de tous les individus du ménage actuellement, quel est le montant mensuel des revenus nets (de contributions sociales et CSG) dont votre ménage dispose ?

2- Si vous ne pouvez donner un montant précis, pouvez-vous au moins en donner une estimation ? (on propose à l'enquêté une liste de 9 tranches de revenu)

L'opinion du ménage sur son niveau de vie :

Si on considère à présent les ressources mensuelles de votre ménage, diriez-vous qu'elles vous permettent de vivre :

1. très difficilement
2. difficilement
3. assez difficilement
4. assez aisément
5. Aisément
6. très aisément

En s'intéressant à la phase précédant l'interview, on cherche en particulier à savoir si l'hypothèse d'homogénéité du comportement de réponse est acceptable ou non.

Sur les seuls ménages répondants, la méthode est appliquée sur la durée de l'interview, et sur deux questions suivantes portant sur le montant du revenu mensuel, et sur l'appréciation du niveau de vie. La déclaration du revenu a donné lieu à deux traitements, l'un portant sur le montant de revenu libellé en chiffres, le deuxième sur le mode de déclaration choisi, en chiffres ou en tranches.

Sur ces questions, l'effet enquêteur sur la non-réponse partielle n'a pas été testée, parce que les non-réponses sont rares.

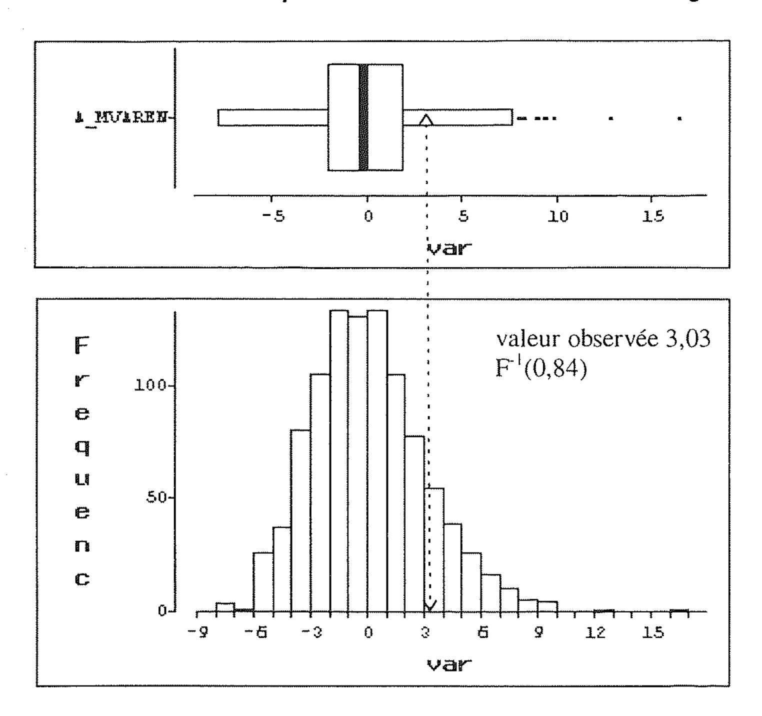
Le test d'un lien entre déclaration de revenu et l'opinion du ménage sur son propre niveau de vie a été construit de la manière suivante : tout d'abord, un revenu par unité de consommation à été calculé en appliquant une échelle d'équivalence (le

poids du premier adulte est 1, le poids de chaque personne âgée de plus de quatorze ans est 0,5, celui de chaque enfant est 0,3). Puis des tranches de revenu ont été déterminées pour présenter les mêmes effectifs par tranche que l'opinion sur le niveau de vie. Ensuite une variable de cohérence a été calculée pour chaque ménage comme la différence entre les niveaux d'échelle du ménage pour le revenu et le niveau de vie.

Pour une variable donnée, deux distributions ont été simulées : celle de la variance enquêteur $\hat{\sigma}^2_{enq}$, et celle de la différence des moyennes des groupes 1 et 2.

Pour la variance enquêteur comme pour la différence entre groupes, si la valeur observée est décalée sur la droite de la distribution simulée, on lit dans ce décalage l'impact de l'effet enquêteur sur la statistique choisie. Par contre, le fait que cette valeur observée se situe vers le centre de la distribution permet d'exclure un effet enquêteur (la distribution simulée est à peu près normale, le centre correspond à la fois à la moyenne, au mode, et à la médiane, donc au fractile 0,5).

Graphique 3 :
Distribution de la variance enquêteur du revenu mensuel du ménage



Les résultats de l'ensemble des simulations réalisées sont résumés dans le tableau ci-dessous par la donnée du fractile auquel correspond la valeur observée, c'est-à-dire celle propre aux deux groupes d'enquêteurs.

Tableau : Fractiles des valeurs observées dans les distributions simulées

Champ retenu	Variable	Fractile* de la variance enquêteur observée	Fractile* de la différence observée
Tous les logements tirés	Statut de l'enquête	0,62	0,63
Tous les logements hormis les hors champ	Non réponse totale	0,64	0,54
Les ménages répondants à l'enquête	Durée de l'interview	1	1

* Lecture : Pour le statut de l'enquête, la variance enquêteur observée se trouve placée dans la distribution simulée de telle sorte que 62% des valeurs de la distribution lui sont inférieures.

Champ retenu	Variable	Fractile* de la variance enquêteur observée	Fractile* de la différence observée
Les ménages donnant leur revenu en chiffres	Revenu mensuel du ménage	0,84	0,91
Les ménages donnant leur revenu en chiffres ou en tranche	Mode de déclaration du revenu	0,99	0,99
Les ménages répondants à la question sur le niveau de vie	Opinion du ménage sur son niveau de vie	0,69	0,48
Les ménages répondants aux questions niveau de vie et revenu	Cohérence de classement entre niveau de vie et revenu	0,75	0,80

En résumé

L'effet enquêteur ne joue ni sur le statut de l'enquête ni sur la non-réponse totale quand on en confond les différentes causes possibles. On légitime ainsi l'hypothèse d'homogénéité des comportements de réponse, qui autorise à modéliser l'effet enquêteur à partir des seuls répondants.

Pour le statut de l'enquête en trois modalités, la moyenne est utilisée comme un résumé de la distribution. La nature qualitative de la variable rendrait préférable un test portant sur une distance entre distributions, plutôt que sur une différence de moyennes. Ce travail n'a pas encore été réalisé.

Avec la même limite portant sur l'utilisation de la moyenne, l'effet enquêteur a été testé sur le statut de l'enquête plus détaillé, en différenciant les différentes causes de non-réponse (ménage impossible à joindre, absent de longue durée, déclaré inapte à répondre, ou refus de répondre). Ce statut détaillé, comme la durée d'interview déclarée, font apparaître un fort effet enquêteur. Ces deux variables ont en commun d'être renseignées directement par l'enquêteur. Elles ne sont que très peu sensibles à une interaction entre l'enquêteur et l'enquêté. Elles sont entachées d'un fort effet individuel de l'enquêteur.

L'effet paraît nul sur l'appréciation du niveau de vie. Il n'est pas très fort sur le montant du revenu. Sur la cohérence de classement entre échelle de revenu et

échelle de satisfaction du niveau de vie, l'effet est un peu plus élevé que sur le seul niveau de vie. Puisque l'effet touche un décalage entre les deux échelles calculé en moyenne par enquêteur, il laisse soupçonner un décalage à tendance un peu systématique par enquêteur.

Quant au mode de déclaration du revenu, il peut être considéré comme déterminé par l'enquêteur.

BIBLIOGRAPHIE

Cochran W. G., *Sampling Techniques*, Wiley.

Deville J.-C.(1994), *Quelques éléments pour l'analyse de l'effet enquêteur dans le dispositif « europanel »,* Note Insee N°780/F401.

Särndal C.-E., Swensson B., Wretman J., *Model Assisted Survey Sampling*, Springer-Verlag.

DES ENQUÊTEURS À LA RENCONTRE DES ENTREPRISES : UNE NOUVELLE APPROCHE

Chantal de Barry, Marcel Perrot

Introduction

Pour la plupart des enquêtes auprès des entreprises, l'envoi postal demeure la méthode initiale de collecte, contrairement aux enquêtes auprès des ménages où s'est imposée dès l'origine la collecte par enquêteurs. Cette pratique s'est d'autant plus confortée que les données d'entreprises se prêtent mal à une collecte directe, exigeant souvent recherche et élaboration préalables. Ainsi dans leur presque totalité, ces enquêtes sont effectuées par voie postale et gérées sur dossier par des équipes centralisées ; elles ne donnent que très occasionnellement lieu à des déplacements.

Toutefois, au cours de la période récente, l'évolution de l'environnement est venue affecter ce fonctionnement traditionnel : l'observation des données et le recueil ont cessé d'apparaître aussi aisés. La réalité des entreprises devient plus complexe, leur contour, leur mode d'organisation dépasse le cadre classique d'une seule entité juridique et rend difficile l'observation statistique. Par ailleurs, les entreprises acceptent de moins en moins la charge de réponse aux enquêtes statistiques qui ne se distinguent par forcément des formulaires administratifs. En plus, elles ne comprennent pas toujours la finalité et l'utilité de nos opérations.

Afin d'améliorer la qualité de l'observation et des relations avec les entreprises, l'Insee entreprend une expérimentation de collecte directe par enquêteurs en complément du mode de collecte traditionnel.

Dans cet article nous présenterons d'abord les pratiques courantes d'enquêtes auprès des entreprises en France et à l'étranger. Ensuite, nous préciserons quelles sont les difficultés qui motivent la collecte directe par entretien. Enfin, nous aborderons l'expérimentation en cours et ses premiers résultats.

1. Les pratiques actuelles

1.1 - Les enquêtes auprès des entreprises à l'Insee

1.1.1 Les caractéristiques

L'Insee envoie chaque année pas moins de 500 000 questionnaires aux entreprises.

Ces enquêtes peuvent se classer en **trois** catégories :

- des enquêtes à périodicité **infra-annuelle** (principalement trimestrielle) à vocation conjoncturelle, portant chacune sur 3 000 à 4 000 entreprises « Conjoncture », « Stocks-produits-et-charges », « Enquête mensuelle commerce et services », « Prix de vente industriels ».

- des enquêtes à périodicité **annuelle** de nature structurelle « Enquête annuelle d'entreprise Commerce et Services » (120 000 entreprises), « Liaisons financières » (17 000 questionnaires)

- des enquêtes à périodicité irrégulière **supérieure à l'année**, comportant entre 20 000 et 30 000 questionnaires, concernant les petites entreprises (« Petites entreprises industrielles », « Entreprises nouvellement créées ») ou portant sur les coûts de la main-d'œuvre et la structure des salaires.

Il faudrait ajouter à cela un certain nombre d'enquêtes **ponctuelles** thématiques, au volume moins important : de quelques centaines à quelques milliers de questionnaires.

1.1.2 Des déplacements limités

Ces enquêtes sont pratiquement toutes effectuées par voie postale et gérées sur dossier par des équipes centralisées. Elles ne donnent que rarement lieu à des déplacements sur le terrain.

Quelques déplacements, seulement, s'effectuent dans le cadre des relances de certaines enquêtes, après les rappels postaux. C'est le cas notamment pour l'Enquête annuelle d'entreprise dans le Commerce et les Services et parfois aussi pour le renouvellement d'échantillon de l'enquête mensuelle Commerce-Services. Dans la majeure partie des cas, ces visites sont faites dans les régions par des enquêteurs-pigistes de l'Insee et elles ne concernent au mieux que 1 000 à 2 000 entreprises.

Les enquêtes thématiques, nationales ou régionales, quant à elles, peuvent recourir plus souvent à des enquêteurs-terrain, mais il s'agit là d'interventions limitées et la plupart du temps assurées par des réseaux d'enquêteurs extérieurs à l'Insee.

Une exception notable doit être signalée. L'enquête **« Prix de vente industriels »**, dans sa phase d'initialisation et de renouvellement, est assurée d'une manière permanente par un réseau d'enquêteurs Insee hautement spécialisés. Neuf enquêteurs-terrain effectuent en moyenne plus de 1 000 visites par an. Ils négocient avec un dirigeant de l'entreprise la liste des produits-témoins qui feront l'objet, par la suite, d'un suivi régulier de prix. Cette collecte régulière est ensuite réalisée par voie postale.

Il faut citer également l'Enquête Annuelle d'Entreprise dans les DOM qui est faite en partie par visites.

1.2 - Les enquêtes sur le terrain, hors Insee

1.2.1 Dans les services statistiques ministériels

Au Ministère de l'Equipement, des enquêteurs-terrain sont utilisés pour des enquêtes-prix concernant le « coût de la construction » et les « prix des travaux d'entretien et d'amélioration des logements». Ce sont tous des enquêteurs-pigistes qui opèrent auprès des maîtres d'ouvrage et des entreprises de construction.

Ces exemples révèlent que c'est surtout dans le cas **d'enquêtes portant sur l'observation des prix** que l'on fait appel à des équipes permanentes d'enquêteurs spécialisés.

Au Ministère de l'Agriculture, on trouve, cependant, un cas particulier qui s'en distingue. Ici, un important réseau d'enquêteurs d'environ 500 pigistes, est utilisé pour effectuer les enquêtes auprès des exploitations agricoles. Ce réseau est en place depuis la fin des années 60 et donne toute satisfaction. Il reçoit un très bon accueil de la part des exploitants et obtient un taux de réponse voisin de 100%. Il est intéressant de noter que dans ce monde particulier qu'est l'agriculture, constitué de petites « entreprises » très dispersées, l'enquête par enquêteur-terrain a été privilégiée. Ceci témoigne d'un effort de rapprochement vers l'enquêté qui donne des résultats.

1.2.2 À l'étranger

Les exemples étrangers montrent surtout l'existence d'équipes d'enquêteurs chargées des contacts et des visites auprès des **grandes entreprises**.

Le Canada, la Nouvelle-Zélande, l'Australie, l'Irlande, les Pays-Bas, l'Argentine utilisent habituellement des équipes d'enquêteurs de terrain pour suivre et enquêter les grandes entreprises, en général de l'ordre de quelques centaines d'unités.

Un exemple intéressant existe aux Etats-Unis où le Bureau des Statistiques du Travail conduit une enquête directement auprès 30 000 établissements avec 160 équipes d'enquêteurs de haut niveau. Il s'agit d'une enquête sur les salaires et autres rémunérations effectuée selon un mode de collecte assistée par ordinateur.

2. Un changement des modes de collecte

Pourquoi est-il apparu nécessaire, à l'Insee, de modifier les pratiques traditionnelles d'enquête auprès des entreprises ? Celles-ci tiennent des comptabilités normalisées ; elles sont assujetties à des obligations légales. Il n'y a de ce fait aucun problème de disponibilité de l'information ou de recueil.

Pourtant, cette apparente simplicité masque une autre réalité et de nouvelles exigences nous obligent à modifier nos habitudes d'enquêtes et à aller vers les entreprises.

2.1 - Les motifs du changement

Les difficultés d'observation et de recueil des informations se sont accentuées dans la période récente.

2.1.1 Des difficultés d'observation

Une grande partie des difficultés rencontrées dans les statistiques d'entreprises provient de l'observation.

Aussi, les concepteurs d'enquêtes se préoccupent-ils depuis longtemps des questions de qualité de l'observation. En particulier, le constat fréquent de divergences entre les différentes sources (sources fiscales, enquêtes annuelles d'entreprise, enquêtes de production,...), pouvant entraîner des écarts importants dans les résultats, vient souligner ce problème.

Mais, plus récemment, la complexification croissante des structures des entreprises a été un sujet soulevé à l'occasion des réflexions engagées autour de la rénovation de l'Enquête Annuelle d'Entreprise.

En effet, celles-ci, connaissent une double évolution. D'une part, elles s'insèrent de plus en plus dans des relations de réseau ou de groupe avec d'autres entreprises ; d'autre part, elles délèguent progressivement à l'extérieur nombre de fonctions qui constituaient leurs fonctions traditionnelles (gestion du personnel, comptabilité, recherche et développement, gestion des investissements, commercialisation,...). Ces entreprises « éclatées » ou « imbriquées » deviennent des objets d'observation plus difficiles : par exemple, les facteurs de production qu'elles utilisent peuvent ne plus être comptabilisés chez elles mais dans une autre entreprise et échappent à l'enquête. On obtient alors, si l'on agrège de telles entreprises avec d'autres plus classiques, sans précaution ni examen, des catégories statistiques hétérogènes, perdant toute signification.

2.1.2 Des réactions négatives de la part des entreprises

Par ailleurs, le rejet de la charge administrative de la part des entreprises devient de plus en plus vif. Les revendications d'allégement visent naturellement les enquêtes statistiques : charges plus visibles, d'utilité moins évidente pour les entreprises même si ce ne sont pas, de très loin, les plus lourdes.

Ces réactions négatives peuvent avoir des incidences non négligeables sur la qualité des réponses fournies :

- L'entreprise peut donner une réponse erronée :
  « Quand je ne sais pas quoi mettre, je mets un peu n'importe quoi »
- L'entreprise peut ne pas répondre :
  « Si on me dit à quoi ça sert, je ferai peut être l'effort de répondre »

Or, les enquêtes auprès des entreprises portent sur des unités présentant des disparités importantes. Dans de nombreux secteurs, les 4 premières entreprises représentent souvent plus de 30% du poids total en termes de chiffres d'affaires ou d'emploi salarié ; elles peuvent même atteindre et parfois dépasser 75%. Aussi, un certain nombre de grandes entreprises sont irremplaçables et leur réponse est absolument indispensable pour la validité des résultats.

2.2 - Des actions nouvelles en direction des entreprises

2.2.1 Améliorer les relations avec les entreprises

À la suite de ces constatations, l'Insee a entrepris des actions pour renverser cette évolution qui s'auto-entretient. De nombreuses initiatives ont été prises ces dernières années afin de rechercher l'amélioration, à la fois, de l'image de la statistique et de la qualité de la collecte.

Plusieurs rapports (CALLIES sur la simplification des enquêtes auprès des entreprises, MOTHE-ALLAIN sur la rationalisation du dispositif statistique public) avaient déjà attiré l'attention sur les efforts qui doivent être menés en direction des entreprises.

Dans ce sens, le Conseil National de l'Information Statistique a renforcé sa mission de concertation sur les enquêtes en créant le Comité du Label, dont la mission est d'examiner leur bonne conformité avant leur lancement. Une des conséquences est de demander systématiquement que tous les questionnaires aient fait l'objet de tests préalables auprès des entreprises.

2.2.2 Aller sur le terrain

Mais, les objectifs d'amélioration de l'image de l'Insee et de la qualité de la collecte ne pourront être véritablement atteints que si **des spécialistes se déplacent sur le terrain.**

En effet, se déplacer sur le terrain apporte un ensemble d'avantages :

- gagner l'adhésion, **d'abord en montrant qu'on agit en professionnel, ensuite en manifestant l'intérêt qu'on attache à la réponse de l'entreprise et enfin en expliquant et persuadant ;**
- contribuer à l'allégement de la charge, **en assurant aide, assistance et explication et en adaptant l'interrogation au système d'informations de l'entreprise ;**
- atteindre le bon interlocuteur, **condition préalable à l'obtention d'une réponse de qualité** ;
- améliorer l'observation, **grâce à la visite sur place et aux différentes discussions et rencontres qu'elle permet.**

2.3 - Des interventions et des missions ciblées

Ainsi, une mission double sera confiée à ces enquêteurs de terrain :
- procéder aux entretiens directs dans les entreprises
- être les représentants de l'Insee auprès de celles-ci
(en leur apportant des informations et en recevant leurs demandes)

Mais, bien entendu, pour des raisons de coût et de temps, ils ne remplaceront pas intégralement le mode de collecte actuel, qui s'effectue principalement par voie postale. Ils viendront en appui, en complément, d'une manière sélective, en prenant en compte des critères d'efficacité. Ils devront travailler en relation et en collaboration avec les gestionnaires des enquêtes.

Ils interviendront, donc, auprès de **populations ciblées d'entreprises** :
- des entreprises nouvelles entrant dans le champ d'une enquête : dans le cadre d'un premier contact avec l'entreprise, le recours à un enquêteur semble intéressant pour présenter les objectifs de l'enquête à laquelle l'entreprise va devoir répondre pour la première fois, et remplir avec elle le questionnaire.
- des entreprises non répondantes : après toutes les procédures classiques de relance, l'enquêteur intervient sur place pour convaincre, présenter les objectifs de l'opération, et collecter les données.
- des grandes entreprises à configuration complexe: analyse sur place des organisations complexes (détermination de l'activité exercée, réseaux,), apport en direct des explications sur le questionnaire et collecte des informations compatibles avec les concepts définis par le statisticien.

.... ou pour des opérations particulières, comme :

- les tests de questionnaire : recueillir auprès des entreprises interrogées leurs critiques et suggestions pour améliorer le questionnaire et le rendre compatible avec la réalité des entreprises
- des enquêtes thématiques ponctuelles à champ limité
- la reprise de contact ou la recherche d'un bon correspondant : renouer des contacts avec une entreprise qui a cessé de répondre, entretenir le suivi des relations.

3. L'expérience menée à l'Insee

Dès le mois de septembre 1997, une opération-pilote a été lancée en interne à l'Insee. Quatre directions régionales participent à cette expérimentation : Nancy, Rouen, Saint-Quentin-en-Yvelines et Toulouse où ont été mises en place de petites équipes spécialisées. Trois des quatre directions régionales avaient déjà une équipe compétente dans le domaine des entreprises. De leur côté, les responsables

d'enquêtes ont accepté de s'associer à cette expérience en nous confiant des opérations variées de taille limitée.

3.1 - Une démarche pragmatique

Expérimenter un nouveau mode de collecte auprès des entreprises et en mesurer l'apport nécessite de se donner les moyens de bien l'analyser en mettant en place, préalablement, un dispositif fiable. L'expérimentation de fonctionnement et le test méthodologique constituent deux étapes étroitement dépendantes.

3.1.1 Première étape: une expérimentation du fonctionnement

La première étape de notre démarche a été **d'initialiser le réseau** avec des opérations de petite taille. Il s'agissait de mettre en relation d'un côté, des équipes formées, compétentes présentant une offre crédible de services et de l'autre des responsables d'enquêtes acceptant un nouveau mode de collecte.

Pour constituer ces équipes qualifiées, il a fallu mettre au point un programme de formation contenant différents volets : un module de formation technique consacré aux aspects comptables, juridiques et sociaux des entreprises et des formations spécifiques aux différentes enquêtes. Entrer en contact avec une entreprise et discuter efficacement avec un de ses représentants sur des problèmes d'organisation, de gestion, de comptabilité exigent à la fois un bon niveau de compétence et d'expérience qu'il faut développer dans ces équipes.

Ces premières opérations nous ont permis également de tester tous les rouages nécessaires au bon fonctionnement d'un réseau d'enquêteurs implantés sur plusieurs régions : élaboration du calendrier d'enquête, définition du plan de charge de chaque direction régionale, organisation des actions de formation et mise au point, avec la participation des directions régionales et des responsables d'enquêtes, de documents standards qui facilitent la gestion des enquêtes et l'élaboration de bilans : fiche de propositions de travail, grille d'entretien, compte rendu d'enquête, ...

3.1.2 Deuxième étape : mesure de l'apport du réseau sur la qualité des résultats

L'expérimentation du dernier trimestre 1997, bien que d'ampleur limitée, est riche d'enseignements quant au bilan qualitatif qu'il nous apporte (cf. §4 ci-dessous). Toutefois, les opérations qui se mettent en place pour 1998 visent de nouveaux objectifs : mieux quantifier l'apport du réseau sur la qualité des résultats d'enquêtes comparé au mode traditionnel de collecte par voie postale, recueillir des éléments de coût et de charge.

Cette analyse sera centrée sur des opérations-types comme les relances d'entreprises non répondantes ou des interrogations d'entreprises complexes pour lesquelles on dispose de résultats d'enquêtes antérieures obtenus par voie postale et portera sur des volumes plus importants.

3.2 - Les interventions de la première étape

3.2.1 Quelques chiffres

Le programme du dernier trimestre 1997 prévoyait une centaine d'enquêtes sur le terrain à répartir sur 15 enquêteurs (3 en Haute-Normandie, 6 en Ile-de-France, 2 en Lorraine, et 4 en Midi-Pyrénées).

Les opérations ont porté sur différents types d'interventions

- entretiens en face à face sur des enquêtes thématiques de petite taille
- tests de questionnaires
- relances d'entreprises récalcitrantes (entretien direct).

Elles ont concerné différents domaines : le commerce, les services et l'industrie et abordé différents aspects des entreprises : l'organisation en réseaux, le domaine comptable, l'informatisation dans les secteurs du bricolage et des activités comptables.

Au total, **5 opérations** ont été conduites:

	Visites	Types
Commerce.		
Bricolage	36	Entretiens
Services.		
Ingénierie	9	Relances
Comptables	20	Tests
Télécommunications	10	Tests
Industrie.		
Epei	30	Tests
Total	**105**	

3.2.2 Description des opérations

♦ ***L'enquête thématique sur les « Réseaux de distribution des articles de bricolage ».***

Cette opération s'est trouvée bien adaptée à la dimension de notre réseau-pilote. Il s'agissait d'interroger en face en face les dirigeants d'une quarantaine de têtes de réseau sur des thèmes stratégiques comme leur mode d'organisation et leur informatisation.

Les enquêteurs étaient chargés de prendre rendez-vous avec un interlocuteur « privilégié » de la centrale d'achat (qui avait reçu préalablement un questionnaire) et d'aller collecter sur place les informations. Les principales difficultés dans cette opération ont été la longueur du questionnaire (durée d'entretien 1H, 1H30), les thèmes abordés et la taille des entreprises interrogées (généralement des groupes).

♦ ***Des tests de questionnaires de l'enquête portant sur les « Changements organisationnels et informatisation dans les activités comptables ».***

20 entreprises de toutes tailles dispersées sur le territoire national relevant du secteur des activités comptables ont été sélectionnées pour cette phase de test et 7 enquêteurs ont été mobilisés pour cette opération. Ce secteur rassemble plusieurs types d'entreprises : les cabinets comptables, soumis à un conseil de l'ordre, et les centres de gestion agréés. Un des objectifs du test était de vérifier que le questionnaire s'adaptait à ces deux catégories d'entreprises, ainsi qu'aux différentes tailles de cabinets comptables. Le test visait également la compréhension des questions ainsi que la disponibilité des informations nécessaires. Il permettait enfin d'évaluer la durée de remplissage du questionnaire. Dans la plupart des cas, c'est le directeur lui-même ou le directeur général qui a participé à l'entretien.

♦ ***Des tests de questionnaires de l'enquête portant sur les « Opérateurs et les fournisseurs de services de télécommunication ».***

10 entreprises classées dans les secteurs « télécommunications nationales », « autres activités de télécommunication » et « activités informatiques » ont été sélectionnées pour cette phase de test et 7 enquêteurs du réseau-pilote y ont participé. Les entreprises rencontrées avaient entre 21 et 4 000 salariés et un chiffre d'affaires compris entre 3 millions et 4 milliards de francs. L'interrogation de petites structures dans ce secteur n'a pas été souhaitée du fait d'un questionnement trop détaillé pour ce type d'unités. Les objectifs de ce test étaient de s'assurer de la compréhension des questions auprès des entreprises interrogées, d'évaluer le temps du remplissage du questionnaire.

♦ ***Des tests de questionnaires de l'« enquête auprès des petites entreprises industrielles » (Epei).***

Les tests s'adressaient à une trentaine de petites entreprises de moins de 20 salariés relevant du champ de l'industrie. Les thèmes abordés dans l'enquête portaient sur les caractéristiques générales de l'entreprise et de son dirigeant, les conditions

d'exploitation et d'organisation de ces unités (degré d'informatisation, effort d'innovation technologique, connaissance de la clientèle).
Les tests visaient une bonne compréhension des questions et l'existence des informations demandées. Ils ont mis en évidence une grande difficulté d'approche de ces petites structures : les chefs d'entreprises refusent de répondre, n'ont pas le temps ou sont souvent absents. Les entretiens se sont déroulés parfois dans des conditions difficiles : dérangements fréquents. Enfin, les informations comptables n'étaient pas toujours disponibles.

♦ ***Des relances auprès de grandes entreprises non répondantes du secteur de l'ingénierie.***
L'enquête-pilote européenne sur les services d'ingénierie et d'études techniques portait en France sur un échantillon stratifié par tranche d'effectifs de 515 entreprises. Il s'agissait d'une enquête non obligatoire. L'enquête interrogeait de façon exhaustive toutes les entreprises de plus de 200 salariés (soit 64 entreprises). Le réseau d'enquêteur est intervenu en dernier recours auprès de certaines unités non répondantes de cette strate à l'issue de deux rappels postaux.

4. Le premier bilan

Nous n'avons fait jusqu'à maintenant qu'une expérience limitée, mais riche d'enseignements et encourageante pour la suite de l'expérimentation qui sera conduite en 1998.

Ces observations confirment les effets attendus de la démarche sur le terrain, tant du point de vue de l'amélioration des relations avec les entreprises que de l'amélioration des observations. Nous pouvons dès maintenant dresser un premier bilan qualitatif de ce premier trimestre de démarrage.

4.1 - Amélioration du taux de réponse

L'enquête « Ingénierie ». L'intervention des enquêteurs a atteint une réussite de 75% en obtenant 9 réponses sur 12. Ce qui a permis de faire passer le taux de réponse de la strate exhaustive (+200 salariés) de 36% à 50%, en apportant un gain de 14 points. Ainsi le meilleur taux de réponse a été obtenu dans cette strate représentant près de la moitié du chiffre d'affaires du secteur ; le taux d'ensemble de cette enquête **non obligatoire** n'atteignant que 39%. Les 9 entreprises répondantes aux enquêteurs représentaient environ 7 à 8% du CA , de la VA et des effectifs, mais surtout **15%** des investissements et **20%** des exportations ; ce qui, étant donné la nature de l'enquête, était capital pour l'analyse des résultats.

L'enquête « Bricolage ». Sur 43 visites réelles à effectuer, les enquêteurs ont obtenu 36 entretiens, soit un taux de succès de 84%. Ils n'ont essuyé que 7 refus, dont 3 sont de véritables échecs, les 4 autres correspondant en fait à des entreprises concernées d'une manière secondaire par l'enquête.

Les enquêteurs ont rapporté que leurs visites avaient permis de récupérer *« des enquêtes vouées au panier »*, soulignant par contre que *« les relances postales ou téléphoniques sont souvent mal perçues par les entreprises ».*

Une entreprise n'aurait pas répondu au questionnaire de l'enquête « Bricolage » qu'elle jugeait trop compliqué, mais a accepté d'y répondre en direct, assistée et guidée par l'enquêteur.

4.2 - Amélioration des observations

Les enquêteurs relèvent l'importance du contact direct dans la fourniture de la bonne réponse.

Ils signalent que :
« les entreprises profitent de la présence de l'enquêteur pour demander des éclaircissements sur les questions et manifestent leur souci de donner la bonne réponse »

et rappellent qu'au contraire :
« lorsque les enquêtes sont réalisées par courrier, les entreprises les retournent dans le meilleur des cas, mais ne posent jamais de questions ».

Ils observent tout particulièrement que :
« lors des tests d'enquêtes sur le terrain, les entreprises s'expriment sur la pertinence ou l'oubli de certaines questions, ce qu'elles n'auraient vraisemblablement jamais écrit ».

Les apports des enquêteurs sur la qualité peuvent s'observer tout spécialement dans deux domaines.

4.2.1 Le repérage des activités

L'enquête « Opérateurs de télécommunications » a bien mis en évidence les difficultés que rencontre une entreprise à se classer dans les catégories statistiques.

Certaines entreprises, classées d'après leurs déclarations à l'Enquête Annuelle d'Entreprises dans le secteur « fournisseurs de services de télécommunications (64.2B) », se sont révélées, après discussion avec l'enquêteur, exercer en fait une

activité différente telle que la construction de réseaux câblés, l'installation de pylônes ou le nettoyage de câbles sous-marins.

En effet, les entreprises ne trouvant pas leurs métiers dans la liste proposée par le questionnaire de l'EAE se positionnent dans les « autres services de télécommunications » ou ventilent leur chiffre d'affaires en fonction de la destination finale des équipements qu'ils fournissent, c'est-à-dire le marché sur lequel elles se placent.

C'est ainsi qu'une entreprise dont l'activité était l'installation : installation de pylônes et de réseaux filaires de télédistribution ventilait, en fait, son chiffre d'affaires entre diverses activités de services : services de réseaux fixes, de réseaux mobiles et de télédiffusion.

Une autre, qui construisait des réseaux de câbles et proposait des études techniques, ventilait son chiffre d'affaires entre télédiffusion et autres services.

La recherche du bon classement dans la nomenclature proposée a exigé des tâtonnements et des explications sur les métiers exercés.

4.2.2 Le périmètre de l'unité interrogée

L'enquête « Bricolage », de son côté, a révélé la nécessité de mettre au point avec l'entreprise interrogée le périmètre de l'unité statistique observée. Des écarts importants séparent les représentations des statisticiens et celles des entreprises qui peuvent avoir une vision trop large (groupe) ou trop étroite (établissement) ou tronquée (domaine d'activité) par rapport à l'unité de l'enquête.

Ainsi, après discussions sur place, les enquêteurs ont été amenés à corriger les premières réponses spontanées de l'entreprise.

Une entreprise qui avait rempli le questionnaire avant l'entretien n'avait pas compris l'étendue géographique du questionnaire (sur toutes les régions).

Une autre entreprise au contraire avait intégré des points de vente vis-à-vis desquels elle n'avait pas de rôle de tête de réseau, mais seulement des participations.

4.3 - Amélioration des relations

Le travail d'information et de communication auprès des entreprises se renforce.

Les visites des enquêteurs dans les entreprises sont généralement l'occasion de donner des informations sur l'Insee et souvent de recevoir des demandes de la part de celles-ci. C'est parfois le début d'une ouverture de relations avec l'Institut. Globalement, l'image de l'Insee en tire toujours bénéfice.

Dans leurs comptes-rendus, les enquêteurs font état de ces différents points positifs. Les extraits suivants en donnent une illustration.

4.3.1 Amélioration de l'image

« beaucoup d'entreprises rencontrées sont satisfaites d'avoir un ***interlocuteur*** *identifié, à l'Insee »*
« la présence d'un enquêteur donne aux entreprises la possibilité de faire des ***remarques****, ou des* ***critiques*** *sur les diverses enquêtes, de pouvoir enfin être entendues, de rappeler le volume d'enquêtes et de relances dont elles font l'objet, sur une même période ou sur des thèmes voisins »*
« le fait qu'un enquêteur Insee se déplace est considéré comme une preuve de ***crédibilité*** *et de* ***sérieux*** *de l'Institut : la présence d'un enquêteur, c'est pour l'entreprise une sorte de vérification des données collectées donc une garantie de fiabilité des résultats »*

4.3.2 Informations sur l'Insee

« les entretiens d'enquêtes se poursuivent au-delà du temps nécessaire et débouchent sur une ***présentation de l'Insee*** *»*
« à chaque enquête il y a toujours une ***présentation de l'Insee*** *»*
« présentation de la ***brochure sur les publications de l'Insee*** *avec description des informations qu'elles contiennent ; ceci suscite toujours de l'intérêt de la part des entreprises »*
« la ***documentation Insee*** *laissée dans l'entreprise lui permet de découvrir parfois l'Insee et souvent les produits Insee dont elle ignorait l'existence »*

4.3.3 Ouverture de relations nouvelles

« une entreprise visitée par un enquêteur a rappelé, par la suite, pour demander des ***informations sur d'autres sujets*** *»*
« une entreprise demande des informations sur le ***recensement*** *»*

5. L'avenir

L'expérimentation se poursuit en 1998, afin d'étendre le test sur une année pleine. La formation des enquêteurs sera approfondie et leur expérience renforcée de façon à les préparer à des interventions plus délicates. La nécessaire couverture du territoire pour limiter les coûts financiers et humains implique des relais régionaux plus nombreux et bien implantés. Le déploiement complet du dispositif et l'ouverture de ce réseau aux services statistiques ministériels devraient s'opérer à l'issue de la phase expérimentale et des conclusions des discussions menées dans le cadre de la démarche sur l'organisation de la production statistique à l'Insee.

5.1 - Une expérimentation plus poussée en 1998

5.1.1 Les interventions prévues pour 1998 sont plus nombreuses

On retrouve, comme en 1997, des actions de relances auprès d'entreprises non répondantes pour l'enquête comptable sur les « Stocks, Produits et Charges », ou l'enquête annuelle dans les secteurs des services, des tests de questionnaire sur l'enquête portant sur les entreprises récemment créées ou des enquêtes thématiques ponctuelles.

Toutefois, quelques opérations nouvelles ont été introduites, comme l'interrogation d'entreprises entrant dans le champ d'une enquête, l'interrogation de grandes entreprises aux structures complexes ou le test du questionnaire de l'enquête annuelle auprès de nouveaux secteurs des services « l'hébergement touristique ».

Au total, le programme d'enquête pour 1998 s'établit à un peu plus de **800 enquêtes sur le terrain**.

5.1.2 De nouveaux objectifs

Disposant désormais d'équipes initiées au domaine des entreprises et à la pratique des entretiens sur le terrain, l'expérimentation menée sur 1998 sera davantage axée sur des investigations méthodologiques. Certaines opérations ont été plus particulièrement sélectionnées pour **mesurer l'apport du réseau d'enquêteur sur l'amélioration des résultats d'enquête.**

Il s'agit de l'enquête « Stocks, Produits et Charges » qui permettra, dans la phase de relance des non répondantes, une comparaison des résultats obtenus après activation des procédures d'imputation, d'une part, et après intégration des données collectées par enquêteur, d'autre part. Avec l'envoi d'enquêteurs sur le terrain auprès des unités récalcitrantes, on prévoit d'améliorer les taux de réponse et la qualité des réponses et d'en faire une évaluation. Ces gains en qualité s'observeront sur quelques variables de cadrage de l'enquête (chiffre d'affaires, valeur ajoutée, excédent brut d'exploitation).

Les grandes entreprises complexes répondant à l'enquête « Stocks, Produits et Charges » feront l'objet également d'une comparaison de résultats entre les données de l'enquête collectées par enquêteur et celles collectées l'année précédente par voie postale. Des mesures de l'écart sur les variables d'accrochage comme les stocks et leur ventilation ainsi qu'une analyse comparée des non-réponses partielles seront réalisées. Ce même test sera effectué sur les questionnaires simplifiés de l'EAE-Commerce.

Sur l'ensemble des opérations conduites durant l'année 1998, nous procéderons également à des **évaluations des coûts et des charges** des enquêtes terrain. Des informations détaillées sur les dépenses et le temps passé à la préparation des entretiens, aux entretiens proprement dits et aux déplacements seront recueillies pour chaque visite d'entreprise ; nous pourrons, ainsi, mieux définir la charge d'enquête par enquêteur, et mieux équilibrer, à l'avenir, le plan de charge des différentes équipes régionales.

Enfin, l'ensemble de ces travaux menés sur 1997 et 1998 avec les concepteurs d'enquête, l'expérience acquise par les équipes d'enquêteurs régionaux sur le terrain devraient nous aider à mettre au point un **guide méthodologique** de collecte à l'usage des enquêteurs.

5.2 - Extension territoriale

L'incomplète couverture du territoire liée à une démarche de test n'est pas toujours compatible avec la réalisation d'enquêtes réelles. Cette configuration gêne les concepteurs d'enquêtes qui doivent limiter, autant que possible, leurs opérations à certaines zones géographiques. Du côté des enquêteurs, les déplacements occasionnés par des enquêtes pour lesquelles on ne peut contrôler la localisation des entreprises, peuvent être importants et peser lourd en terme de coûts humains et de coûts financiers.

La réflexion menée sur l'organisation de la production statistique à l'Insee a intégré la mise en place d'un réseau d'enquêteurs couvrant l'ensemble du territoire. L'extension du dispositif à un plus grand nombre de directions régionales interviendra progressivement suite à cette phase expérimentale.

5.3 - Ouverture vers les services statistiques ministériels

Des possibilités d'offre de services vers les SSM pourraient être envisagées dans l'avenir. Le service statistique du ministère du travail est déjà intéressé par l'utilisation du réseau d'enquêteurs de l'Insee. Celui-ci souhaiterait disposer d'enquêteurs pour relancer et convaincre, par téléphone, les nouvelles entreprises lors du renouvellement de l'échantillon de l'enquête trimestrielle sur l'Activité et les

Conditions d'Emploi de la Main-d'Œuvre (Acemo). Ces interventions exigent de bien connaître les entreprises et de savoir argumenter. Il s'agit d'une enquête auprès des établissements, donc bien adaptée au champ d'action du réseau régional.

A terme, ces enquêteurs représentants du système statistique public auprès des entreprises pourraient jouer un rôle central dans la coordination et l'harmonisation du système d'enquêtes en devenant progressivement **les correspondants uniques** des entreprises pour les différentes enquêtes. Ils contribueraient à en donner une image plus cohérente.

BIBLIOGRAPHIE

Rapport CALLIES-TAILHADES - « Réseau d'enquêteurs-entreprises » n°105/B005 - Janvier 1995

Rapport de BARRY-PERROT - « Réseau d'enquêteurs auprès des entreprises » - n°272/E210 - Octobre 1997

E. VERGEAU, N. CHABANAS - « Le nombre de groupes d'entreprises a explosé en 15 ans » - Insee Première n°441 - Novembre 1997

« Réseau de distribution, relations producteurs-distributeurs, EDI, le cas du bricolage » - Lettre du SSE n°25 - Octobre 1997

« Des enquêteurs à la rencontre des entreprises » - Lettre du SSE n°26 - Décembre 1997

« Les unités statistiques au service d'une meilleure représentation de l'économie » - Lettre du SSE n°27 - Février 1998

« Business survey methods » - 1995 - Edité par B.G. Cox, D.A. Binder, B.N. Chinnapa, *et alii* ...

« Observer et représenter un monde de plus en plus complexe : un défi pour la statistique d'entreprises » - Lettre du SSE n°15 - Décembre 1995

Guillaume CASTERA - « Les relations de la division Prix de vente industriels avec les entreprises » - Juin 1997

Documents internes :

Compte-rendu des tests menés sur l'enquête auprès des opérateurs et fournisseurs des services de télécommunication - note n°18/E420 du 21 janvier 1998

Compte-rendu des tests menés sur l'enquête sur les changements organisationnels et l'informatisation dans les activités comptables - Note n°276/E420 du 31 décembre 1997

Compte-rendu de tests menés sur l'enquête auprès des petites entreprises industrielles - note n°18/E210 du 28 janvier 1998

Compte-rendu de tests menés sur l'enquête « Commercialisation des articles de bricolage » - note n°232/E414 du 27 novembre 1997

LA CARTOGRAPHIE INFRACOMMUNALE DE L'INSEE

Philippe Houssay

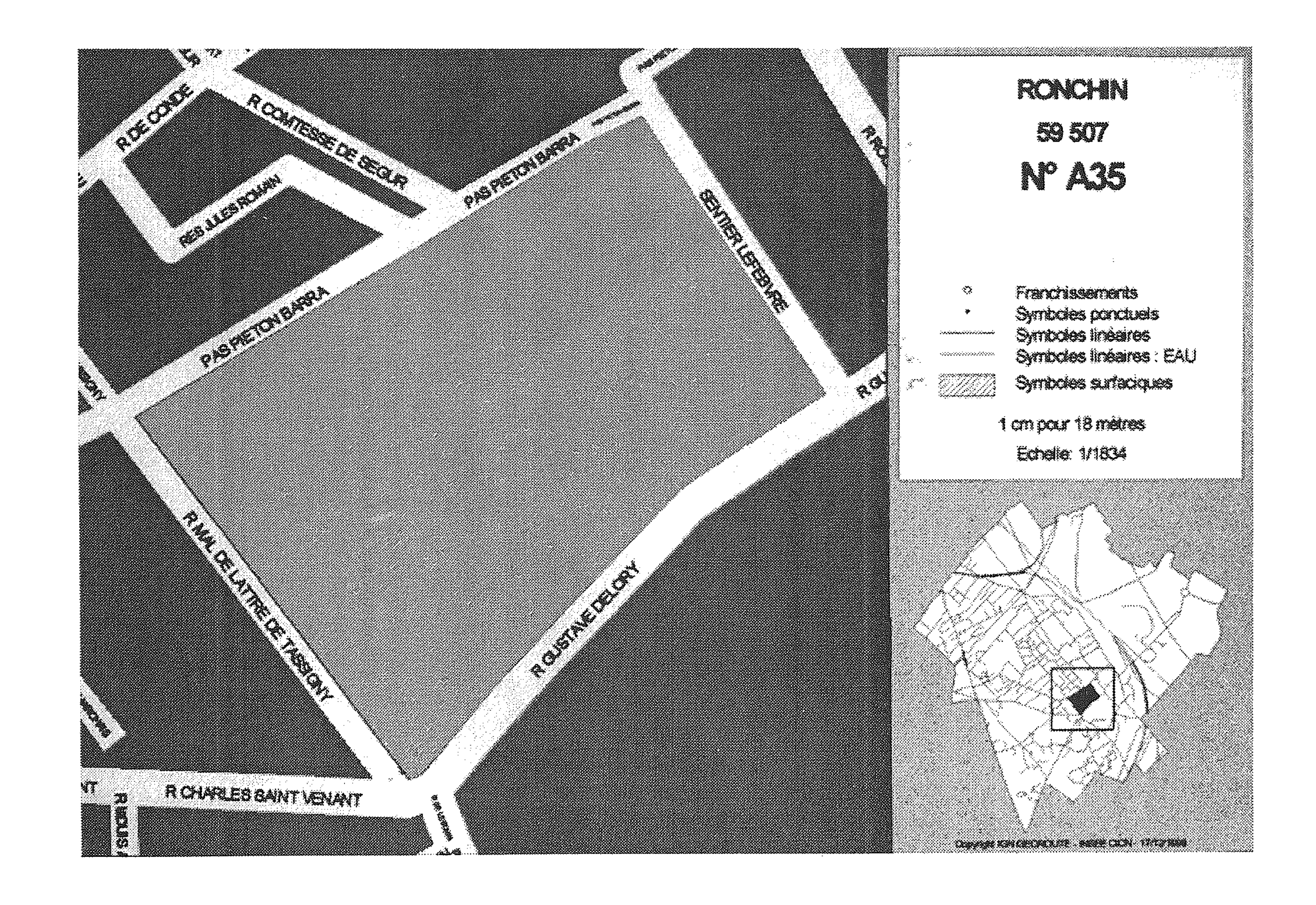
Depuis quelques années l'Insee a considéré que les outils géographiques et cartographiques sont essentiels pour ses activités statistiques. L'apparition sur le marché de logiciels de gestion des SIG (systèmes d'information géographiques) et la puissance des microordinateurs ouvrent des possibilités insoupçonnées il y a encore 10 ans.

L'examen de trois applications typiques permet d'appréhender l'intérêt de disposer d'une cartographie infracommunale numérisée telle que nous l'avons conçue.

1 - La collecte du recensement

Un des premiers besoins à satisfaire a été de réaliser automatiquement la cartographie des districts de recensement destinés aux agents recenseurs. Au cours du projet CICN (cartographie infracommunale numérisée), un test a montré que la représentation des bâtiments sur les plans n'était pas indispensable.

Exemple (voir page suivante)

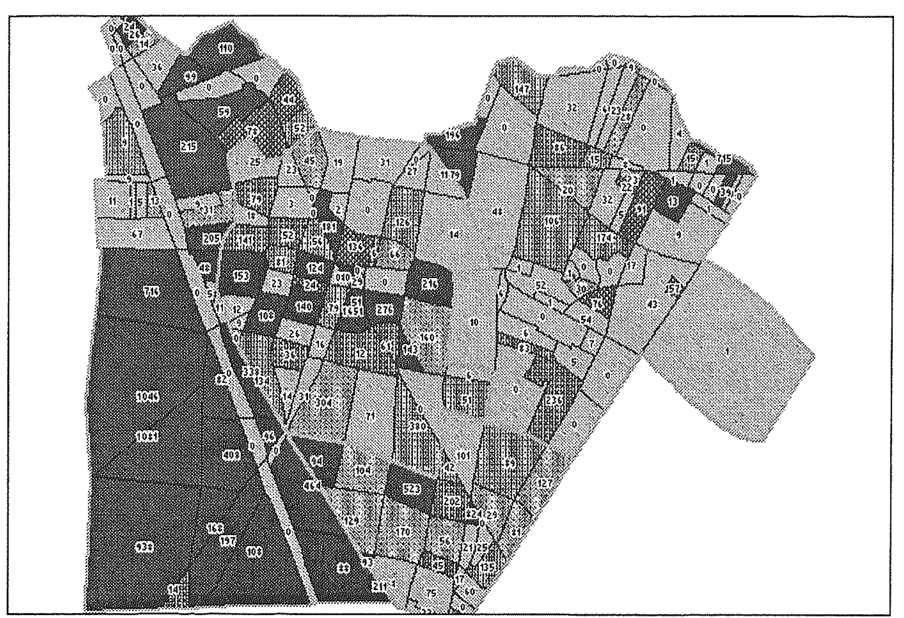

R DE CONDE
R COMTESSE DE SEGUR
RES JULES ROHAN
PAS PIETON BARRA
PAS PIETON BARRA
SENTIER LEFEBVRE
R MAL DE LATTRE DE TASSIGNY
R GUSTAVE DELORY
R CHARLES SAINT VENANT
RONCHIN
59 507
N° A35
Franchissements
Symboles ponctuels
Symboles linéaires
Symboles linéaires : EAU
Symboles surfaciques
1 cm pour 18 mètres
Echelle: 1/1834

2 - La cartographie thématique à l'îlot

Même si la publication dans le grand public de statistiques à l'îlot n'est pas autorisée par la CNIL (Commission Nationale Informatique et Libertés), elle est possible dans le cadre d'études pour certaines missions de service public, comme la politique de la ville.

Les SIG sont un outil puissant pour cela, à condition de disposer d'une cartographie numérisée.

Exemple : Commune d'Aubervilliers, nombre de salariés des établissements de chaque îlot et ratio salariés/habitants.



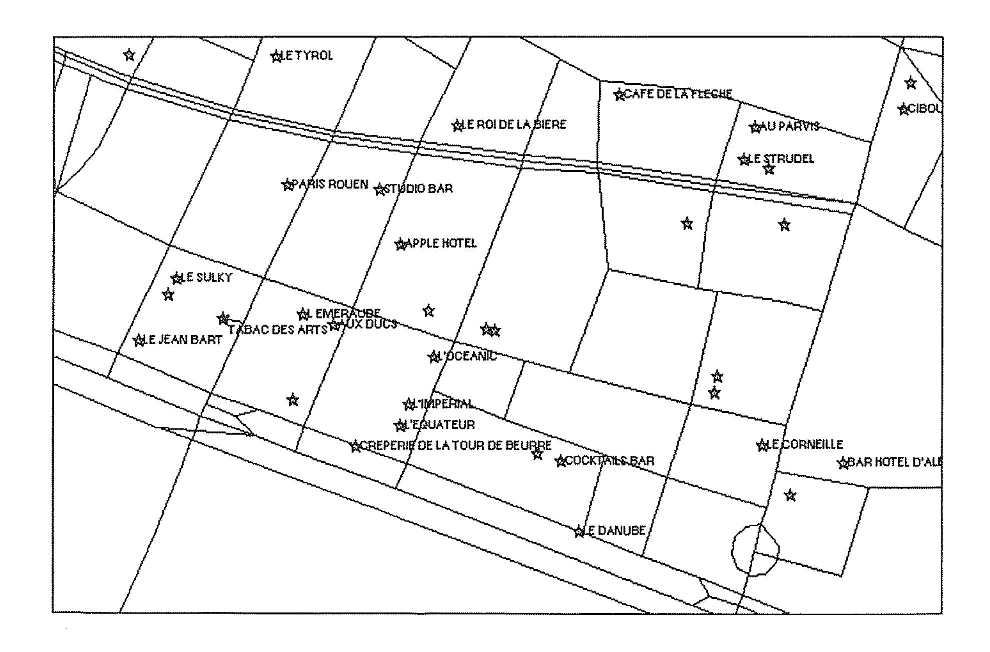
3 - Le géocodage à l'adresse

Dès les années 80 l'Insee a utilisé le découpage du territoire en îlot pour agréger des données statistiques localisées par une adresse dans les fichiers administratifs (exemple : allocataires des services sociaux, salariés des établissements).

Les infrastructures géographiques utilisées par l'Insee permettent une localisation plus précise, en x - y.

L'exemple suivant montre la position exacte des cafés-hôtels-restaurants dans le centre de Rouen.

Exemple :



Ces trois exemples montrent bien que le produit Base-îlots issu de CICN n'est pas une simple cartographie numérisée, mais qu'il s'agit plutôt d'une base de données géographiques à usage multiples.

Le projet CICN (cartographie infracommunale numérisée), commencé en 1992, avait en fait huit objectifs, et on voit bien qu'ils ne concernent pas seulement la collecte du recensement et des enquêtes qui font l'objet des présentes journées de méthodologie :

a) collecte du RP99

b) tirage d'aires, plans pour enquêtes ménage, logement, etc.

c) diffusion: repérage des îlots et diffusion de leur délimitation vers l'extérieur (plans papier ou images)

d) diffusion: définition de zonages à la demande

e) études ou production: géocodage à 10m près et îlotage

f) étude et diffusion: cartographie thématique à l'îlot

g) études par analyse spatiale: croisement de données statistiques à l'îlot avec des données géocodées à l'adresse et des objets géographiques (réseaux, POS, transport, etc.)

h) diffusion de Base-îlots et Replic en tant que tels (permettant ainsi aux clients des usages tels que cités plus haut). Source de recettes et facilite la diffusion de statistiques localisées sous forme de bases de données.

Au vu de tous ces objectifs, on voit qu'il est très réducteur d'assimiler le produit Base-îlot à une simple cartographie numérisée. Il s'agit plutôt d'une base de données géographiques à usages multiples.

4 - L'expérience de la Réunion

En 1997 la direction régionale de l'Insee de la Réunion avait à son programme une enquête de base sur 20 000 ménages, une enquête logement et une enquête famille.

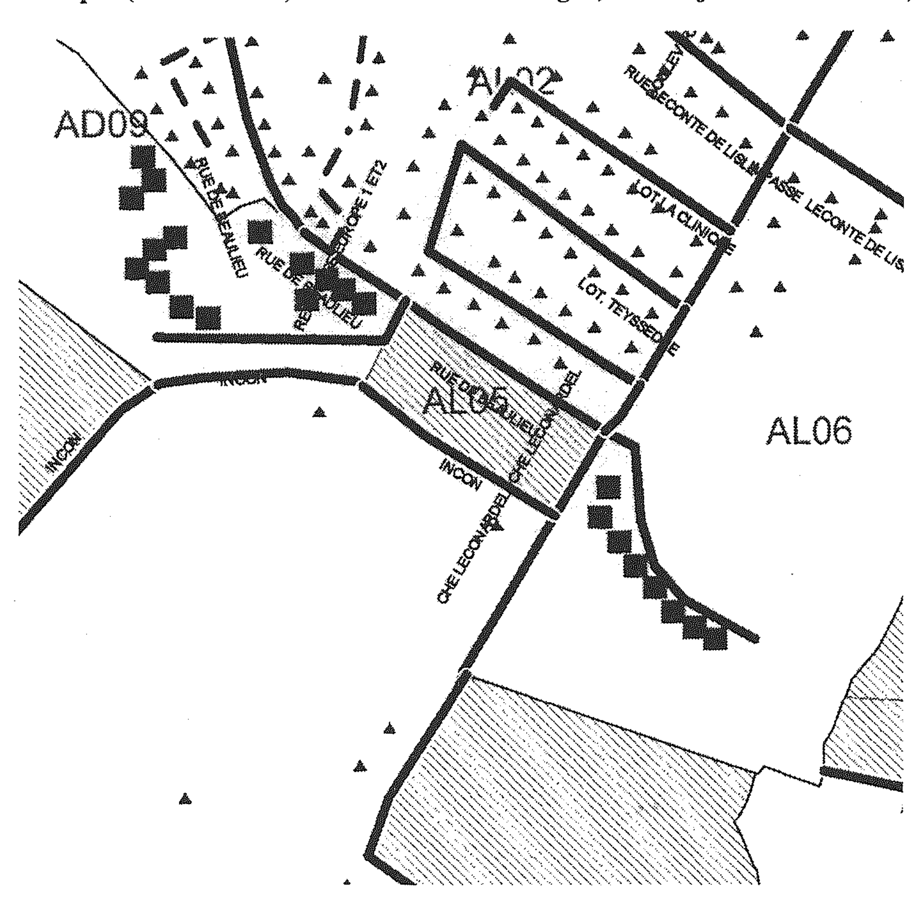
Un processus innovant a été mis en place :

1 - une cartographie numérique des voies et des districts obtenue par **numérisation** de plans (conforme à Base-îlots mais sans adresses),

2 - sur un tiers des districts : un levé sur le terrain de la position relative des immeubles (avec quelques caractéristiques comme le nombre de logements) et de la mise à jour des voies,

Cette opération a été réalisée par des enquêteurs de l'Insee munis de micros portables sans clavier (« pencomputers »). Un logiciel -**PRAO (préenquête assistée par ordinateur)**-a été développé pour leur permettre d'afficher le plan sur leur écran, de corriger les voies et d'ajouter les entrées d'immeubles et leurs variables associées.

Exemple: (les immeubles, ici des carrés et des triangles, ont été ajoutés sur le terrain)



3 - le rapatriement de ces informations sur un SIG (système d'information géographiques), permettant d'assurer la cohérence géométrique, d'archiver les données et d'éditer des plans de travail (voir 5).

4 - le tirage d'aire contiguës de 10 logements environ sur des critères de proximité et d'appartenance à un district,

5 - l'édition de cartes destinées aux enquêteurs décrivant les aires retenues.

Conclusion de l'expérience

- L'utilisation du logiciel PRAO a été bien perçu par les préenquêteurs.

- Le fonctionnement du SIG a été maîtrisé par le personnel de la Direction Régionale.

- Les plans des aires de sondage ont été un outil efficace pour les enquêteurs.

L'expérience de la Réunion a montré qu'il était possible de bâtir un Répertoire d'Immeubles Localisé et de l'utiliser comme base de sondage.

L'expérience a paru suffisamment concluante pour que la Direction Régionale propose d'utiliser le même processus (création d'un RIL par levé sur le terrain à l'aide de PRAO) pour la collecte du RP 99.

Contenu de Base-îlots (référentiel obtenu par le projet CICN) comparé à d'autres sources cartographiques

	Base-îlots (Insee)	PCI	BDTopo	Orthophoto numérique
géométrie des voies	axes des voies (filaire)	emprise du domaine public	filaire	image
libellés des voies	attribut des voies	visuels	non	non
adresses	adresses aux extrémités des tronçons	visuelles	non	non
bâti	non (projet RIL)	contours du bâti cadastral	contours du bâti (photogrammétrie)	image
îlot	contours construits sur filaire	non	non	non
autre	éléments de repérage	parcelles et divers	carte topographique (relief, occupation du sol, etc.)	image
mise à jour	1 à 2 ans (à normaliser)	problématique	5 ans	nouvelle prise de vue

Session 3

Le logiciel de calcul de précision POULPE

LE LOGICIEL POULPE : ASPECTS MÉTHODOLOGIQUES

Nathalie Caron

Sommaire

Introduction

Le logiciel POULPE (Programme Optimal et Universel pour la Livraison de la Précision des Enquêtes) écrit en langage macro SAS, permet d'évaluer la précision de statistiques issues d'enquêtes par sondage complexes, en particulier les enquêtes auprès des ménages ou des entreprises réalisées par l'Insee. Son utilisation suppose de pouvoir décrire rigoureusement le plan de sondage et de disposer des données permettant de calculer les probabilités d'inclusion.

Outre le traitement des données issues de plans de sondage classiques comportant un nombre arbitraire de degrés de tirage, de strates et des procédures de tirage diverses (probabilités inégales, tirages systématiques,...), le logiciel POULPE intègre aussi le traitement des enquêtes en plusieurs phases, la prise en compte de la correction de la non-réponse ainsi que celle du redressement par le logiciel CALMAR.

Le type de statistique dont le logiciel POULPE est capable de chiffrer la variabilité est général. Ainsi, la statistique peut être le total d'une variable ou une statistique complexe, fonction de plusieurs totaux de variables (moyennes, ratios,...). Dans ce dernier cas, la méthode de linéarisation, programmée dans le logiciel, permet de se ramener à l'estimation de la variance d'un total d'une variable synthétique.

Ce papier constitue une version courte d'un document plus ambitieux qui recense l'ensemble des principaux éléments méthodologiques utilisés ou développés par l'UMS (Unité Méthodes Statistiques) pour mettre au point la première version du logiciel. Ce document paraîtra prochainement dans la série « Méthodologie Statistique » des documents de travail de l'Insee.

Par rapport aux autres logiciels présents sur le marché permettant d'évaluer la précision des enquêtes par sondage, le logiciel POULPE présente les spécificités suivantes :

- des formules d'estimation de la variance utilisables dans le cas d'un sondage « direct » : sondage aléatoire simple à probabilités égales, sondage systématique, sondage équilibré, sondage aléatoire simple à probabilités inégales. Dans ce dernier cas, on utilise une formule d'estimation de variance qui ne nécessite pas la connaissance des probabilités d'inclusion double ;

- *l'utilisation de la récursivité* : les formules récursives d'estimation de variance dans les plans de sondage complexes permettent d'estimer la variance d'un estimateur en appliquant successivement les formules d'estimation de variance

dans le cas d'un sondage « direct » aux différents « branches » de l'arbre décrivant le plan de sondage ;

- *le traitement des enquêtes en deux phases* dans le cas où la première phase est quelconque et la seconde phase est un sondage poissonnien ou un sondage stratifié. Cette particularité du logiciel est très importante pour l'INSEE ; en effet, beaucoup d'enquêtes ménages ont un plan de sondage à deux phases avec une sur ou sous représentation d'une partie des logements au moment du tirage. De plus, en assimilant le traitement de la non-réponse totale par repondération à une phase supplémentaire, cette spécificité nous permet de prendre en compte la correction de la non-réponse ;

- *le traitement des enquêtes en trois phases* dans le cas où la première phase est quelconque, la seconde phase est un sondage stratifié et la troisième est un sondage poissonnien.

1ère partie - Traitement des sondages « directs »

On appelle sondage « direct », les algorithmes qui extraient d'un fichier un échantillon ayant certaines propriétés. Ce terme s'oppose à celui de « sondage décomposé » où l'échantillonnage est décomposé en plusieurs sous-échantillonnages. Dans un ultime sous-échantillonnage (c'est-à-dire dans le sous-échantillonnage conduisant à la sélection des unités enquêtées), on réalise un sondage « direct ». La sélection d'unités d'échantillonnage non ultimes se fait également par ce type de sondage.

Après avoir défini les notations utilisées, nous décrirons les différents types de sondage « directs » disponibles dans le logiciel.

I. Notations

La variable d'intérêt est notée ***Y*** et nous souhaitons estimer son total $Y = \sum_1^N Y_i$.

On notera $\pi_i = P(i \in \text{échantillon})$, $\pi_{ij} = P(i \text{ et } j \in \text{échantillon})$, les probabilités d'inclusion simple et double.

L'estimateur classique d'Horvitz Thompson défini par $\hat{Y}_\pi = \sum_{i \in s} \frac{y_i}{\pi_i}$ est un estimateur sans biais de Y (c'est-à-dire que la moyenne pondérée des valeurs de cet estimateur obtenues sur tous les échantillons possibles de taille n correspond à la vraie valeur Y).

Sa variance et son estimation de variance sont respectivement :

$$V(\hat{Y}_\pi) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \quad \text{et}$$

$$\hat{V}(\hat{Y}_\pi) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

II. Principaux types de sondage directs

Trois principaux types de sondage directs sont disponibles dans le logiciel POULPE : le sondage aléatoire simple, le sondage systématique et le sondage à probabilités inégales.

Sondage aléatoire simple

Dans le cas d'un sondage aléatoire simple sans remise, les formules de variance et d'estimation de variance se simplifient et deviennent :

$$V\left(\hat{Y}_\pi\right) = N^2(1-f)\frac{S^2}{n} \text{ où } S^2 = \frac{\sum_{i\in U}\left(y_i - \overline{Y}\right)^2}{N-1}$$

$$\hat{V}\left(\hat{Y}_\pi\right) = N^2(1-f)\frac{s^2}{n} \text{ où } s^2 = \frac{\sum_{i\in s}\left(y_i - \overline{y}\right)^2}{n-1} \text{ et } f = \frac{n}{N}.$$

Sondage systématique

D'après la théorie de l'échantillonnage, comme un sondage systématique est la réalisation d'un sondage en grappe où l'on ne sélectionne qu'une seule grappe, il est impossible d'estimer la variance pour ce type de tirage. Cependant, sous diverses hypothèses liées à la modélisation, il est possible d'obtenir des estimateurs corrects de la variance. Ainsi, en supposant que les données sont rangées dans le même ordre que celui précédant le tirage systématique, la formule retenue dans le logiciel POULPE est :

$$\hat{V}\left(\hat{Y}_\pi\right) = N^2(1-f)\frac{t^2}{n}$$

où $t^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}\left(y_i - y_{i+1}\right)^2$ avec y_i qui représente la valeur de la variable Y pour le ième individu du fichier .

Sondage à probabilités inégales

Le cas des tirages à probabilités inégales pose un problème beaucoup plus délicat que les types de sondage directs présentés ci-dessus. En effet, les formules

d'estimations de variance ne se simplifient pas et par conséquent présentent deux inconvénients :

- le premier est de recourir à des sommes doubles qui sont numériquement lourdes à calculer et sans doute assez instables. Ainsi, lorsque l'échantillon compte 100 individus, les sommes doubles comptent environ 5 000 termes. Cette difficulté est toutefois surmontable, surtout si on réalise qu'en pratique, on ne mettra en œuvre le tirage à probabilités inégales que pour des échantillons qui n'excèdent pas une ou deux dizaines d'unités ;

- le second inconvénient est plus délicat à résoudre. Les formules en question font en effet appel aux probabilités d'inclusion d'ordre 2 $(\text{les } \pi_{ij})$, probabilités pour que les couples (i, j) soient dans l'échantillon. Or, sauf dans le cas de sondage à probabilités égales, les probabilités doubles π_{ij} ne sont pas connues et ne sont pas calculables pratiquement ou au prix de calculs importants.

Dans le cadre des sondages à probabilités inégales, plusieurs formules approximatives d'estimation de variance se ramenant à des sommes de carrés et ne faisant pas intervenir les probabilités d'inclusion doubles ont été comparées :

① La première formule d'approximation de la variance est due à B. ROSEN (1991)

$$\hat{V}ar_{Rosen}\left(\hat{Y}\right)=\frac{n}{n-1}\sum_{s}\left(\frac{y_k}{\pi_k}-D\left(\frac{y}{\pi}\right)\right)^2\left(1-\pi_k\right)$$

avec $$D\left(\frac{y}{\pi}\right)=\sum_{s} a_k\frac{y_k}{\pi_k}\Big/\sum_{s} a_k$$

$$a_k=\left(1-\pi_k\right)\log\left(1-\pi_k\right)\ /\pi_k$$

② Deux autres approximations peuvent être obtenues à partir de la théorie décrite par J.-C. DEVILLE (1993) :

$$\hat{V}_1\left(\hat{Y}_\pi\right)=\frac{n}{n-1}\sum_{s}\left(1-\pi_i\right)\left(\frac{y_i}{\pi_i}-D_2\left(\frac{y}{\pi}\right)\right)^2$$

D'où $D_2\left(y/\pi\right)$ est la moyenne pondérée par les $\left(1-\pi_i\right)$ des quantités $\frac{y_i}{\pi_i}$.

$$\hat{V}_2\left(\hat{Y}_\pi\right) = \frac{1}{1 - \sum_{i=1}^{s} a_i^2} \sum_{i \in s} (1 - \pi_i)\left(\frac{y_i}{\pi_i} - D_2\left(\frac{y}{\pi}\right)\right)^2$$

avec $a_i = (1 - \pi_i) \Big/ \sum_{i \in s} (1 - \pi_i)$.

Les simulations décrites dans le document de J.-C. DEVILLE et C. VITE SAN-PEDRO (1993) indiquent que quelle que soit la taille de la population, les résultats sont satisfaisants dès que la taille de l'échantillon dépasse 8. Par contre, pour de très petits échantillons et de très petites populations, on obtient une sous-estimation de la variance pouvant atteindre 20%.

La formule retenue dans le cadre du logiciel est $\hat{V}_1\left(\hat{Y}_\pi\right)$.

2ème partie - Traitement des sondages décomposés

I. La stratification

La population est scindée en H parties (appelées strates) à partir d'informations auxiliaires. On réalise un tirage indépendamment dans chacune de ces strates. Un estimateur sans biais du total de la variable Y est $\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h$ où $\hat{Y}_h$ est un estimateur du total de la variable au sein de la strate h. Les tirages étant indépendants d'une strate à une autre, on a $V(\hat{Y}) = \sum_{h=1}^{H} V(\hat{Y}_h)$

II. Le tirage à plusieurs degrés

Dans le cas de sondage à plusieurs degrés, on utilise un système récursif dû à J. DURBIN (1955) et amélioré par D. RAJ (1966) puis par J.N.K. RAO (1975). Nous nous contenterons dans ce document d'en exposer le principe général. Plaçons-nous tout d'abord dans le cas d'un sondage en grappe. L'estimateur du total de la variable d'intérêt est : $\hat{Y} = \sum_{k \in s} y_k / \pi_k$ où s est l'échantillon des grappes et π_k la probabilité d'inclusion d'ordre 1 de la grappe k. Pour les trois types de sondage exposés dans la partie précédente, on connaît un estimateur de la variance $f(y_s)$, où $f(y_s)$ désigne une forme quadratique des y_k de s. Dans le cas d'un sondage aléatoire simple, par exemple, on a en notant N la taille de la population, n celle de l'échantillon et $\overline{y}$ la moyenne dans l'échantillon :

$$f(y_s) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{s} (y_k - \overline{y})^2 / (n-1)$$

Plaçons-nous maintenant dans le cas d'un sondage à deux degrés. Dans ce cas, le total de la variable Y dans l'unité primaire k noté y_k est remplacé par un estimateur $\hat{y}_k$ bâti sur un échantillon s_k d'unités secondaires de l'unité primaire considérée. Un estimateur du total de la variable d'intérêt Y est donc :

$$\hat{y}_2 = \sum_{k \in s} \hat{y}_k / \pi_k \, .$$

D'après RAJ (1966), un estimateur de variance sans biais pour cet estimateur est

$$\hat{V}\left(\hat{y}_2\right) = f\left(\hat{y}_s\right) + \sum_s \frac{v_k}{\pi_k}.$$

où v_k est un estimateur sans biais de la variance de $\hat{y}_k$ et $f\left(y_s\right)$ est l'estimateur de variance correspondant au premier degré.

Ce principe, exposé dans le cas d'un plan de sondage à deux degrés, se généralise à un plan de sondage à n degrés à condition de le décrire par un arbre dont les nœuds sont l'Univers, ***les unités primaires, les unités secondaires*** ... A chaque nœud est associé un type de sondage noté TS (Sondage aléatoire simple à probabilités égales, inégales,...) indiquant la façon dont sont échantillonnées les données au niveau du nœud considéré (se reporter à l'article de J.-N. Petit).

Pour obtenir une estimation de la variance, il faut disposer pour chaque TS mis en œuvre dans le plan de sondage :

- d'une formule donnant un estimateur sans biais de la forme :

$$\hat{t} = \sum \frac{y_k}{\pi_k} \qquad \text{(linéaire)}$$

- d'une formule donnant un estimateur de la variance de $\hat{t}$:

$$f^{TS}(\ldots y_k \ldots) \qquad \text{(quadratique)}$$

On "remonte" l'arbre depuis les éléments terminaux jusqu'à l'Univers. Ainsi dans tout plan de sondage complexe, la variance d'un total estimé par l'estimateur d'Horvitz-Thompson peut être évaluée à partir des fonctions f relatives à chaque forme de sondage direct et de procédés récursifs qui viennent d'être décrits.

III. Les sondages en plusieurs phases

Nous détaillons dans cette partie le traitement dans POULPE des enquêtes en deux phases. Celui des enquêtes en trois phases (correspondant à un plan de sondage en deux phases dont la seconde phase est elle-même composée de deux phases) repose sur le même principe.

III.1. Principe

Un sondage en 2 phases correspond à 2 sondages successifs qui s'appliquent aux mêmes unités :

① un premier échantillon s_1 est sélectionné à l'aide des techniques proposées précédemment. On suppose qu'une information supplémentaire est disponible pour toutes les unités de s_1 ;

② un échantillon s_2 est alors tiré de s_1 en utilisant les techniques précédentes.

La réalisation d'une enquête en deux phases permet en particulier de recueillir une information auxiliaire auprès de l'échantillon 1ère phase et de l'utiliser pour optimiser le tirage de l'échantillon deuxième phase. De plus, la correction de la non-réponse globale est généralement appréhendée par une modélisation d'une phase de sondage supplémentaire. Ainsi, les enquêtes de l'Insee se modélisent en général en trois phases : la seconde phase correspond à une sur-représentation au moment du tirage de l'échantillon et la troisième correspond à une correction de la non-réponse totale.

III.2. Sondages en 2 phases

III.2.1. Cas général

1ère phase :
Un échantillon s_1, de taille n_1, est tiré d'une population U selon un plan de sondage PS_1. On note π_i la probabilité d'inclusion de l'unité i , π_{ij} la probabilité d'inclusion double des unités i et j et $\Delta^1_{ij} = \pi_{ij} - \pi_i \, \pi_j$.

2ème phase :
Un échantillon s_2, de taille n_2, est tiré de s_1 selon un plan de sondage PS_2. On note les "probabilités d'inclusion" liées à ce tirage : p_i, p_{ij} et $\Delta^2_{ij} = p_{ij} - p_i \, p_j$.

Estimateurs
On montre que (voir par exemple C.-E. SÄRNDAL, B. SWENSSON et J. WRETMAN (1992), p. 347 et suivantes) :

- $\hat{Y} = \sum_{i \in s_2} \frac{y_i}{\pi_i p_i}$ est un estimateur sans biais du total $Y = \sum_{i \in U} Y_i$

$\cdot \hat{Y} = \sum_{i \in s_2} \frac{y_i}{\pi_i p_i}$ est un estimateur sans biais du total $Y = \sum_{i \in U} Y_i$

$$\cdot V(\hat{Y}) = \sum_{i \in U} \sum_{j \in U} \Delta^1_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} + E_{s_1}\left[\sum_{i \in s_1} \sum_{j \in s_1} \Delta^2_{ij} \frac{y_i}{\pi_i \, p_i} \frac{y_j}{\pi_j \, p_j} \right]$$

= variance 1ère phase + variance 2ème phase

où E_{s_1} désigne l'espérance relativement à la loi de probabilité de s_1.

$$\cdot \hat{V}(\hat{Y}) = \sum_{i \in s_2} \sum_{j \in s_2} \frac{\Delta^1_{ij}}{\pi_{ij} \, p_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} + \sum_{i \in s_2} \sum_{j \in s_2} \frac{\Delta^2_{ij}}{p_{ij}} \frac{y_i}{\pi_i \, p_i} \frac{y_j}{\pi_j \, p_j} \quad (1)$$

= variance estimée 1ère phase + variance estimée 2ème phase

$\hat{V}(\hat{Y})$ est un estimateur sans biais de $V(\hat{Y})$.

Programmation de la formule (1) dans le logiciel

- Sous réserve d'une description complète du plan de sondage PS_1, le logiciel sait calculer des quantités de la forme :

$$\sum_{i \in s_1} \sum_{j \in s_1} \frac{\Delta^1_{ij}}{\pi_{ij} \, \pi_i \, \pi_j} z_i \, z_j = \sum_{i \in s_1} \sum_{j \in s_1} A_{ij} \, z_i \, z_j$$

$$\text{avec} \quad A_{ij} = \frac{\pi_{ij} - \pi_i \, \pi_j}{\pi_{ij} \, \pi_i \, \pi_j}, \; A_{ii} = \frac{1 - \pi_i}{\pi_i^2}$$

Cette quantité correspond à la variance estimée du total de la variable Z, pour le sondage en une phase PS_1. On rappelle que les ***termes*** A_{ij} ***ne sont en général pas explicites***, mais calculés récursivement et implicitement dans le logiciel (car les π_{ij} ne sont pas connues), à l'aide de formules, exactes ou approchées, permettant d'estimer la variance à chaque degré de tirage.

- De même, sous réserve d'une description de PS_2, le logiciel sait calculer des quantités de la forme :

$$\sum_{i \in s_2} \sum_{j \in s_2} \frac{\Delta_{ij}^2}{p_{ij}\, p_i\, p_j} z_i\, z_j = \sum_{i \in s_2} \sum_{j \in s_2} B_{ij}\, z_i\, z_j$$

$$\text{avec } B_{ij} = \frac{p_{ij} - p_i\, p_j}{p_{ij}\, p_i\, p_j} \quad , \quad B_{ii} = \frac{1 - p_i}{p_i^2}$$

C'est donc la 1ère partie de la formule (1) qui pose problème, car elle s'écrit :

$$\sum_{i \in s_2} \sum_{j \in s_2} \frac{A_{ij}}{p_{ij}} y_i\, y_j \quad \text{ou encore : } \sum_{i \in s_1} \sum_{j \in s_1} \frac{A_{ij}}{p_{ij}} z_i\, z_j \qquad (1')$$

$$\text{en posant } z_i = \begin{cases} y_i \text{ si } i \in s_2 \\ 0 \text{ sinon} \end{cases}$$

Même si les p_{ij} étaient connues (ce qui n'est pas le cas dès que PS_2 est un peu complexe), l'expression (1') n'est pas calculable telle quelle par le logiciel, ni en dehors du logiciel puisque les A_{ij} $(i \neq j)$ **ne sont en général pas connues**.

J.C. DEVILLE (1993) a proposé des "éléments pour une solution générale" de ce problème. La complexité de la programmation à mettre en œuvre dans ce cadre a conduit à écarter temporairement cette solution, étant donné que les cas de sondage en 2 phases rencontrés dans la pratique conduisent à des simplifications dans la formule (1). Ces "cas simples" sont obtenus lorsque le sondage 2ème phase est :

- poissonnien (c'est-à-dire que chaque unité i de l'échantillon 1ère phase est sélectionnée de façon indépendante avec une probabilité connue p_i), ce qui va permettre la prise en compte par le logiciel de certains traitements de la non-réponse

- stratifié, avec un sondage aléatoire simple dans chaque strate, ce qui va permettre de traiter les enquêtes en 2 phases tirées dans l'échantillon-maître ainsi que le traitement de la non-réponse par groupes de réponse homogène.

Dans ces deux cas, le principe consiste à décomposer le terme (1) en différents termes qui peuvent être calculés soit par récursivité par le logiciel soit directement car ils ne font intervenir que des probabilités d'inclusion simples. Nous examinons successivement ces deux exemples.

III.2.2. Sondage 2ème phase Poissonnien

Un plan de sondage poissonnien conduit à des probabilités d'inclusion doubles qui vérifient :

$$p_{ij} = p_i \; p_j \text{ si } i \neq j \Rightarrow \Delta^2_{ij} = 0 \quad \text{si } i \neq j$$

Avec cette hypothèse, la formule (1) s'écrit :

$$\hat{V}(\hat{Y}) = \sum_{i \in s_2} \sum_{j \in s_2} \frac{A_{ij}}{p_{ij}} y_i \, y_j + \sum_{i \in s_2} \frac{p_i (1-p_i)}{p_i} \left(\frac{y_i}{\pi_i \, p_i} \right)^2$$

$$= \sum_{i \in s_2} \sum_{j \in s_2} \frac{A_{ij}}{p_i \, p_j} y_i \, y_j + \sum_{i \in s_2} A_{ii} \, y_i^2 \left(\frac{1}{p_i} - \frac{1}{p_i^2} \right) + \sum_{i \in s_2} \frac{1-p_i}{p_i^2} \frac{y_i^2}{\pi_i^2}$$

$$= \text{(1a)} + \text{(1b)} + \text{(2)}$$

$$\text{(1a)} = \sum_{i \in s_1} \sum_{j \in s_1} A_{ij} \, y_i^* \, y_j^*$$

où y^* est une nouvelle variable définie sur s_1 par : $y_i^* = \begin{cases} \dfrac{y_i}{p_i} & \text{si } i \in s_2 \\ 0 \text{ sinon} \end{cases}$

Cette quantité se calcule dans la "fonction arbre" du logiciel.

(1b) et (2) qui sont des sommes simples se calculent directement à partir du fichier de données d'enquête.

La somme des deux termes 1a et 1b correspond à l'estimation de la variance 1ère phase. Le terme 2 correspond à l'estimation de la variance seconde phase. Le logiciel POULPE génère automatiquement la variable y^* à partir de la variable d'intérêt *Y* et calcule les trois termes.

III.2.3. Sondage 2ème phase = sondage aléatoire simple stratifié

L'échantillon 1ère phase s_1 est scindé en H strates s_{1h} $(h = 1 \cdots H)$, de tailles N_h. Dans chaque strate s_{1h} on réalise un sondage aléatoire simple sans remise avec le taux de sondage f_h, ce qui détermine un échantillon s_{2h} de taille $n_h = f_h \, N_h$. L'échantillon 2ème phase s_2 est formé par la réunion des s_{2h}.

La quasi totalité des enquêtes en 2 phases tirées dans l'échantillon-maître sont réalisées selon cette méthode, à des fins de sur-représentation de certaines catégories de logements ou de ménages (par exemple sur-représentation de certaines catégories socioprofessionnelles ou de certaines catégories de logements comme dans l'enquête "Conditions de Vie").

Probabilités d'inclusion

$$p_i = f_h \quad \text{si} \quad i \in s_{1h}$$

$$\begin{cases} p_{ii} = f_h & \text{si} \quad i \in s_{1h} \\ p_{ij} = f_h \; f_{h'} & \text{si} \quad i \in s_{1h} \text{ et } j \in s_{1h'} \; , h \neq h' \\ p_{ij} = f_h \dfrac{n_h - 1}{N_h - 1} = f_h^2 \Big/ \left(1 + \dfrac{1 - f_h}{n_h - 1} \right) & \text{si} \quad i \text{ et } j \in s_{1h} \; , i \neq j \end{cases}$$

Le premier type de probabilité d'inclusion double correspond au cas où les unités i et j appartiennent à des strates de seconde phase différentes ; le second au cas où i et j appartiennent à la même strate.

Estimateurs

$$\hat{Y} = \sum_h \sum_{i \in s_{2h}} \frac{y_i}{\pi_i \; f_h} = \sum_h N_h \left[\frac{1}{n_h} \sum_{i \in s_{2h}} \underbrace{\frac{y_i}{\pi_i}}_{= \ell_i} \right] = \sum_h N_h \; \bar{\ell}_h$$

La variance estimée de cet estimateur s'écrit :

$$\hat{V}(\hat{Y}) = \sum_{i \in s_2} \sum_{j \in s_2} \frac{A_{ij}}{p_{ij}} y_i \; y_j + \text{variance estimée pour un SAS stratifié de la variable } \frac{y_i}{\pi_i} = \ell_i$$

$$= \sum_{i \in s_2} \sum_{j \in s_2} \frac{A_{ij}}{p_{ij}} y_i \; y_j + \sum_h N_h^2 \frac{1 - f_h}{n_h} \left(\frac{1}{n_h - 1} \sum_{i \in s_{2h}} (\ell_i - \bar{\ell}_h)^2 \right)$$

$$\textcircled{1} + \textcircled{2}$$

Calcul dans le logiciel

$$\left.\begin{aligned} ① &= \sum_{i\in s_2}\sum_{j\in s_2} A_{ij}\frac{y_i}{p_i}\frac{y_j}{p_j} \\ &+\sum_h \sum_{i\in s_{2h}}\sum_{j\in s_{2h}} A_{ij}\frac{y_i}{p_i}\frac{y_j}{p_j}\frac{1-f_h}{n_h-1}\end{aligned}\right\} \text{(1a)}$$

$$+\sum_h \sum_{i\in s_{2h}} A_{ii}\, y_i^2 \frac{n_h-N_h}{f_h(n_h-1)} \qquad \text{(1b)}$$

$$\text{(1a)} = \sum_{i\in s_1}\sum_{j\in s_1} A_{ij}\, z_i^0\, z_j^0 + \sum_h \sum_{i\in s_1}\sum_{j\in s_1} A_{ij}\, z_i^h\, z_j^h$$

où $z^0, z^1 \ldots z^h \ldots z^H$ sont des nouvelles variables **définies** sur s_1 par :

$$z_i^0 = \begin{cases} \dfrac{y_i}{p_i} & \text{si } i \in s_2 \\ 0 & \text{sinon} \end{cases}$$

$$z_i^h = \begin{cases} \dfrac{y_i}{p_i}\sqrt{\dfrac{1-f_h}{n_h-1}} & \text{si } i \in s_{2h} \quad h = 1\ldots H. \\ 0 & \text{sinon} \end{cases}$$

(1b) et (2) se calculent directement à partir du fichier de données d'enquête.

La somme des deux termes 1a et 1b correspond à l'estimation de la variance 1ère phase. Le terme 2 correspond à l'estimation de la variance seconde phase. Le logiciel POULPE génère automatiquement les variables $z^0, z^1 \ldots z^h \ldots z^H$ à partir de la variable d'intérêt *Y* et calcule les trois termes dont seul le terme 1a s'obtient avec le principe de récursivité.

III.3. Remarques

D'après plusieurs études menées à partir de l'enquête "Conditions de Vie" réalisée par l'Insee en 1993, qui est une enquête en deux phases sur-représentant au moment du tirage les logements "défavorisés" (N. CARON (1996)), nous avons obtenu des résultats surprenants. En effet, les variances des estimateurs de paramètres d'intérêt à

faible coefficient de variation sont estimées à partir du logiciel par des quantités fortement négatives ou anormalement faibles.

Ce phénomène s'explique par le fait que pour les enquêtes en deux phases liées à une sur-représentation au niveau du tirage de l'échantillon dont la seconde phase a été réalisée par un tirage systématique tout en conservant l'ordre du tirage (c'est-à-dire unité primaire après unité primaire), nous nous plaçons dans le cas le plus défavorable au sens où l'estimation de la variance première phase est négative. En multipliant le nombre d'unités primaires, le phénomène ne fait que s'accentuer comme c'est le cas sur l'exemple de l'enquête "Situations Défavorisées". En effet, le plan de sondage de 1ère phase de cette enquête est à plusieurs degrés et son plan de sondage de 2de phase est stratifié avec réalisation d'un tirage systématique dans chaque strate. Ainsi, les variances estimées de la première phase par le logiciel pour l'estimation du nombre de femmes ainsi que pour celle du nombre d'actifs sont respectivement **$-17\ 10^9$ et $-3\ 10^9$.**

Dans la version actuelle du logiciel, la solution au problème des estimations de variance négatives est la suivante : si le logiciel POULPE a obtenu une estimation de variance (de la première phase) négative, l'utilisateur en est averti par un message.

La solution de remplacement qui sera développée dans une version ultérieure du logiciel sera la méthode de réplication. Celle-ci permettra de reconstituer artificiellement un échantillon première phase et d'évaluer sur ce dernier l'estimation de variance due à la première phase.

IV. Le calcul du Design Effect

Le logiciel POULPE calcule la précision obtenue en considérant que les données sont issues d'un plan de sondage aléatoire simple. Ce calcul permet de comparer la variance obtenue par le plan de sondage complexe à celle qu'on aurait obtenue si le plan de sondage avait été celui d'un plan de sondage aléatoire simple. Le rapport des deux estimations de variance est appelé en théorie de sondage le « design effect » (deff) ; il permet en particulier d'apprécier un effet grappe si le plan de sondage est à plusieurs degrés.

IV.1. Définition

Soit $\hat{Y}$ $\left(\text{respectivement } \hat{Y}_{SAS}\right)$ l'estimateur d'un total Y d'une variable *Y* pour un plan de sondage quelconque (respectivement un plan de sondage aléatoire simple sans remise de taille fixe égale à celle de l'échantillon effectivement disponible) ;

Par définition,

$$\text{Deff} = \frac{V(\hat{Y})}{V_{SAS}(\hat{Y}_{SAS})}$$

où $V(\hat{Y})$ est la variance de $\hat{Y}$ et $V_{SAS}(\hat{Y}_{SAS})$ est la variance de $\hat{Y}_{SAS}$ en considérant un plan de sondage aléatoire simple.

IV.2. Estimation de l'effet de sondage pour les enquêtes en une phase

Le "Design Effect" est estimé par :

$$\widehat{\text{Deff}} = \frac{\hat{V}(\hat{Y})}{\hat{V}_{SAS}(\hat{Y}_{SAS})}$$

où $\hat{V}(\hat{Y})$ est estimé à partir du logiciel POULPE avec le "vrai" plan de sondage et où $\hat{V}_{SAS}(\hat{Y}_{SAS})$ est un estimateur sans biais de $V_{SAS}(\hat{Y}_{SAS})$. On démontre qu'un « bon » estimateur de $V_{SAS}(\hat{Y}_{SAS})$ sous un plan de sondage complexe est :

$$\hat{V}_{SAS}(\hat{Y}_{SAS}) = \left[\frac{1}{n}\left(1-\frac{n}{N}\right)\left\{N\sum_{s}\frac{y_k}{\pi_k} - \left(\sum_{s}\frac{y_k}{\pi_k}\right)^2 + \hat{V}(\hat{Y})\right\}\right]$$

qui peut différer nettement de l'estimateur traditionnel $N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}$ si les poids sont très différents.

Or :

$$\hat{V}(\hat{Y}) = \widehat{\text{Deff}}\ \hat{V}_{SAS}(\hat{Y}_{SAS})$$

En approximant N par $\hat{N} = \sum_{s} \frac{1}{\pi_k}$, et $1 - \frac{\widehat{Deff}}{n}\left(1 - \frac{n-1}{N-1}\right)$ par 1, nous obtenons :

$$\widehat{Deff} \approx \frac{\hat{V}\left(\hat{Y}\right)}{\frac{1}{n}\left(1 - \frac{n-1}{\hat{N}-1}\right)\hat{N}\sum_{s}\frac{1}{\pi_k}\left(y_k - \bar{y}\right)^2}$$

avec : $\bar{y} = \sum_{s}\frac{y_k}{\pi_k} \Big/ \sum_{s}\frac{1}{\pi_k}$

IV.3. Estimation de l'effet de sondage pour les enquêtes en plusieurs phases.

Pour les enquêtes en plusieurs phases, $\hat{V}_{SAS}\left(\hat{Y}_{SAS}\right)$ est estimée par :

$$\frac{1}{r}\left(1 - \frac{r-1}{\hat{N}-1}\right)\hat{N}\sum_{s}\frac{1}{\pi_k^*}\left(y_k - \bar{y}\right)^2$$

où r est le nombre d'individus dans la dernière phase et π_k^* représente la probabilité d'inclusion "totale" de l'unité k (c'est-à-dire le produit des probabilités d'inclusion qui correspondent aux différentes phases).

3ème partie - Traitement des statistiques complexes

On appelle statistique complexe toute fonction non linéaire de totaux de variables, comme un ratio, c'est-à-dire le rapport de deux totaux. Pour le traitement de leur estimateur, le logiciel procède comme le logiciel de calcul de la précision suédois CLAN par linéarisation des fonctions à estimer, à l'aide de la formule de Taylor de développement en série (approche formalisée par Woodruff en 1971). Ainsi, le principe consiste à remplacer le calcul de la précision d'une statistique complexe, par celui d'un estimateur d'un total d'une variable artificielle qui est une fonction linéaire des variables observées dont on sait estimer la variance. Ce principe est détaillé ci dessous.

I. Notations

On s'intéresse à l'estimation d'une fonction de q totaux sur la population définie par :

$\theta = f\left(Y_1 \dots Y_k \dots Y_q\right)$ où $Y_k = \sum_{i \in U} y_{ki}$ = total de la variable Y_k

On note $\hat{Y}_k = \sum_{i \in s} \frac{y_{ki}}{\pi_i}$ l'estimateur de Horvitz-Thompson de Y_k

Dès que f est une fonction non linéaire, le paramètre θ est dit "complexe". Celui-ci est estimé en remplaçant chaque total Y_k par son estimateur $\hat{Y}_k$: on obtient ainsi l'**estimateur par substitution** $\hat{\theta} = f\left(\hat{Y}_1 \dots \hat{Y}_k \dots \hat{Y}_q\right)$.

II. Linéarisation de $\hat{\theta}$

On suppose que f est une fonction dérivable, à dérivées partielles continues, et que $\hat{Y}_k - Y_k$ est une "petite" variation (en $O_P\left(1/\sqrt{n}\right)$).

On peut alors écrire :

$$\hat{\theta} - \theta = \sum_{k=1}^{q} \left(\hat{Y}_k - Y_k\right) \frac{\partial f}{\partial Y_k}\left(Y_1 \dots Y_k \dots Y_q\right) + O_p(1/n) = \sum_{k=1}^{q} a_k \left(\hat{Y}_k - Y_k\right) + O_p(1/n)$$

avec $a_k = \frac{\partial f}{\partial Y_k}\left(Y_1 \dots Y_k \dots Y_q\right)$

Le 1er terme de l'expression de droite est d'espérance nulle : $\hat{\theta}$ est donc un estimateur "approximativement sans biais" de θ (le biais est négligeable si n est assez grand).

On peut donc confondre variance et erreur quadratique moyenne, et écrire :

$$V(\hat{\theta}) \approx EQM(\hat{\theta}) = E(\hat{\theta}-\theta)^2 \approx E\left[\sum_{k=1}^{q} a_k\left(\hat{Y}_k - Y_k\right)\right]^2 = V\left(\sum_{k=1}^{q} a_k \hat{Y}_k\right) \qquad \text{puisque } E\hat{Y}_k = Y_k$$

$$= V\left[\sum_{k=1}^{q} a_k\left(\sum_{i\in s}\frac{y_{ki}}{\pi_i}\right)\right] = V\left[\sum_{i\in s}\left(\frac{1}{\pi_i}\sum_{k=1}^{q} a_k y_{ki}\right)\right] = V\left(\sum_{i\in s}\frac{z_i}{\pi_i}\right)$$

en introduisant la **variable linéarisée Z** définie par :

$$z_i = \sum_{k=1}^{q}\frac{\partial f}{\partial Y_k}\left(Y_1 \ldots Y_k \ldots Y_q\right)\; y_{ki}$$

On obtient donc : $V(\hat{\theta}) \approx \sum_{i\in U}\sum_{j\in U}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{z_i}{\pi_i}\frac{z_j}{\pi_j}$

qui, si les a_k étaient connus, serait estimée par :

$$\hat{V}(\hat{\theta}) \approx \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{z_i}{\pi_i}\frac{z_j}{\pi_j}$$

Les a_k dépendent des totaux Y_k inconnus ; on remplace ces totaux par leurs estimateurs $\hat{Y}_k$, soit :

$$\hat{a}_k = \frac{\partial f}{\partial Y_k}\left(\hat{Y}_1 \ldots \hat{Y}_k \ldots \hat{Y}_q\right)$$

et on pose $\hat{z}_i = \sum_{k=1}^{q}\hat{a}_k y_{ki}$, ce qui conduit à l'estimation de $V(\hat{\theta})$:

$$\hat{V}(\hat{\theta}) = \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{\hat{z}_i}{\pi_i}\frac{\hat{z}_j}{\pi_j}$$

III. Applications

- *ratio*

$$R = \frac{Y}{X} = f(X, Y), \qquad \hat{R} = \frac{\hat{Y}}{\hat{X}} = f(\hat{X}, \hat{Y})$$

$$\frac{\partial f}{\partial X}(\hat{X}, \hat{Y}) = -\frac{\hat{Y}}{\hat{X}^2} = -\frac{\hat{R}}{\hat{X}}, \qquad \frac{\partial f}{\partial Y}(\hat{X}, \hat{Y}) = \frac{1}{\hat{X}} \Rightarrow \hat{z}_i = \frac{1}{\hat{X}}(y_k - \hat{R}x_k)$$

- *moyenne* = cas particulier d'un ratio, où X est la variable constante égale à 1

$$\hat{z}_i = \frac{1}{\hat{N}}\left(y_k - \hat{\bar{Y}}\right)$$

4ème partie - La prise en compte de la non-réponse totale

Les enquêtes par sondage sont généralement confrontées au problème de la non-réponse. On distingue deux grands types de non-réponse: la **non-réponse totale** lorsqu'un individu échantillonné ne fournit aucune réponse à l'ensemble du questionnaire et la **non-réponse partielle** lorsqu'un individu échantillonné ne répond pas à une partie plus ou moins importante du questionnaire. Chaque type de non-réponse nécessite une technique particulière de correction. Les méthodes de repondération, principalement utilisées pour compenser la non-réponse totale, consistent à augmenter judicieusement le poids d'échantillonnage des répondants. Par contre, dans les méthodes d'imputation employées pour la non-réponse partielle, les réponses manquantes sont remplacées par une (ou plusieurs) valeur(s) "plausible(s)".

Dans la première version du logiciel, seule la correction de la non-réponse totale est prise en compte pour l'évaluation de la variance. L'idée consiste à modéliser le mécanisme de la non-réponse et à introduire une phase de sondage supplémentaire.

D'après la théorie développée par J.C. Deville dans le polycopié du cours pour les statisticiens européens (1997), les techniques usuelles de correction de la non-réponse totale peuvent être traduites par des équations de calage particulières. Réciproquement, l'utilisation de CALMAR sur un fichier dont la non-réponse n'a pas été traitée par une méthode classique de repondération permet non seulement de limiter les fluctuations dues à l'échantillonnage mais aussi de corriger la non-réponse totale. Nous parlerons dans ce cas de non-réponse corrigée implicitement par CALMAR. Ce cas sera développé dans la 5ème partie.

I. Traitement de la non-réponse dans une enquête en une phase au niveau de l'échantillonnage

Comme la présence de non-réponse est modélisée comme une phase de sondage supplémentaire, on se retrouve dans le contexte d'un ***sondage en 2 phases*** :

- ***1ère phase*** : tirage de l'échantillon total s (répondants + non-répondants) $\left(s_1 = s = r \bigcup \bar{r}\right)$
- ***2ème phase*** : tirage de l'échantillon des répondants $\left(s_2 = r\right)$, selon un "plan de sondage" défini par les probabilités de réponse.

p_i = probabilité de réponse de l'unité i

p_{ij} = probabilité pour que les unités i et j répondent.

En général, les comportements de réponse d'unités différentes sont indépendants, c'est-à-dire $p_{ij} = p_i \, p_j$ si $i \neq j$. On se trouve donc dans le cadre d'une enquête en deux phases, à cette nuance près que les probabilités de réponse p_i ne sont pas connues. L'estimation de ces probabilités par des quantités $\hat{p}_i$ peut être obtenue selon différentes méthodes qui introduisent des contraintes supplémentaires (autrement dit, la relation $p_{ij} = p_i \, p_j$ si $i \neq j$ n'est plus forcément valable au niveau des estimations) :

☆ modélisation de la forme $p_i = G\left(z_i^{'} c\right)$, en particulier à l'aide d'un modèle logit où z_i est un vecteur de variables explicatives du comportement de réponse. Les estimations de p_i sont $\hat{p}_i = G\left(z_i^{'} \hat{c}\right)$ où $\hat{c}$ est un estimateur convergent de c.

☆ **groupes de réponses homogènes** : La population est divisée en sous-populations supposées homogènes au sens de la non-réponse. Elles sont constituées après réalisation de l'enquête en examinant en général le critère répond / ne répond pas en fonction de variables connues pour les répondants et les non-répondants. On peut par exemple réaliser une régression logistique sur l'échantillon pour choisir les variables auxiliaires les plus explicatives de la non-réponse ainsi que pour effectuer des regroupements adéquats de modalités pour définir les sous-populations. Le taux de réponse est estimé par le rapport $f_h = \frac{r_h}{n_h}$ où r_h est le nombre de répondants dans la sous-population h et n_h le nombre d'individus de l'échantillon de la sous-population h. Ce mode d'estimation introduit des contraintes supplémentaires qui conduisent à utiliser les formules de la seconde partie en considérant que l'on a réalisé un **SAS stratifié pour la seconde phase.**

☆ redressement direct par CALMAR sur l'échantillon des répondants. Cette méthode est détaillée dans la 5ème partie.

Il suffit donc de remplacer dans les formules de la seconde partie les p_i par les $\hat{p}_i$.

II. Traitement de la non-réponse dans une enquête en deux phases au niveau de l'échantillonnage

On se place dans le cadre d'une enquête en 2 phases, où le sondage 2ème phase est un SAS stratifié, et où la non-réponse totale a été corrigée par une méthode de repondération (autre que CALMAR). La non-réponse génère une 3ème phase de sondage. Les formules à appliquer correspondent à celles d'une enquête en trois phases en estimant les probabilités d'inclusion de la 3ème phase.

5ème partie - La prise en compte du redressement par CALMAR

I. Estimateurs en présence de redressement par calage

On ne considère ici que le cas d'un calage "simple" de l'échantillon "final" sur des totaux connus sur l'ensemble de la population. CALMAR transforme les poids "initiaux" des unités, notés d_i, en poids "finaux", notés w_i, tels que les w_i soient les plus proches des d_i tout en vérifiant les "équations de calage" :

$$\sum_{i \in s} w_i \, x_i = X$$

où

$$\begin{cases} x_i \text{ est un vecteur de variables auxiliaires } \left(x_i^1 \ldots x_i^k\right) \\ X \text{ est le vecteur \underline{connu} des totaux de ces variables sur la population} \end{cases}$$

Le rapport des poids $\frac{w_i}{d_i}$, appelé aussi facteur de calage, est de la forme :

$$\frac{w_i}{d_i} = F\left(x_i' \, b\right) = g_i$$

$$\text{où} \begin{cases} F \text{ est une fonction dépendant de la méthode de calage utilisée} \\ b \text{ est un vecteur de multiplicateurs de Lagrange} \end{cases}$$

L'estimateur par calage du total Y d'une variable d'intérêt est :

$$\hat{Y}_W = \sum_{i \in s} w_i Y_i$$

Selon les cas, les poids initiaux d_i sont :

- les poids de sondage $\frac{1}{\pi_i}$, où les π_i sont les probabilités d'inclusion

- les poids de sondage corrigés de la non réponse $\frac{1}{\pi_i \, p_i}$

D'après l'article de J.-C. DEVILLE et de C.-E. SÄRNDAL (1992), l'estimateur par calage est équivalent (asymptotiquement) à l'estimateur par régression, dont la variance vaut :

$$V\left(\sum_{i \in s} g_i \frac{u_i}{\pi_i}\right)$$

où u_i est le "vrai" résidu de la régression de Y sur les variables de calage, sur U.

Pour estimer la variance de $\hat{Y}_W$, il suffit donc de disposer d'une formule donnant la variance estimée du total d'une variable d'intérêt *Y*, de remplacer les y_i par les $g_i \hat{u}_i$ dans cette formule, où les $\hat{u}_i$ sont les résidus de la régression, *dans s*, de Y sur les variables de calage (où les unités i sont pondérés par les w_i). Les variables résidus sont calculées directement dans le logiciel POULPE.

Une autre variance estimée possible consiste à faire $g_i = 1$. Même si les auteurs recommandent plutôt la première formule, c'est la seconde qui est programmée dans le logiciel puisqu'elle ne requiert pas en entrée les pondérations d'extrapolation finales.

II. Estimation de variance en l'absence de non-réponse

En l'absence de non-réponse (ou en présence de non-réponse mais non corrigée, ni explicitement, ni implicitement), l'application des résultats du paragraphe précédent donne :

II.1. Sondage en une phase

$$\hat{V}\left(\hat{Y}_w\right) = \sum_{i \in s} \sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}\, \pi_i\, \pi_j} \left(g_i\, \hat{u}_i\right)\left(g_j\, \hat{u}_j\right)$$

$$\text{où } g_i = w_i\, \pi_i$$

II.2. Sondage en 2 phases, avec sondage 2ème phase SAS stratifié

On applique la formule donnée dans la partie 2 en remplaçant les y_i par les $g_i\ \hat{u}_i$, où $g_i = w_i\ \pi_i$.

III. Estimation de variance en présence de non-réponse corrigée explicitement

CALMAR transforme les poids $d_i = \frac{1}{\pi_i \hat{p}_i}$ (poids de sondage corrigés de la non-réponse, où les $\hat{p}_i$ sont les probabilités de réponse estimées) en w_i. On peut appliquer les résultats précédents aux cas suivants.

III.1. Sondage en une phase

On applique les formules de la partie 2, en remplaçant les y_i par les $g_i \, \hat{u}_i$, et les p_i par les $\hat{p}_i$.

III.2. Sondage en 2 phases, avec sondage 2ème phase SAS stratifié

On applique les formules d'une enquête en trois phases en remplaçant les y_i par les $g_i \, \hat{u}_i$ et en estimant les probabilités d'inclusion de la 2ème et de la 3ème phase.

IV. Estimation de variance en présence de non-réponse corrigée implicitement par CALMAR

CALMAR transforme les poids $d_i = \frac{1}{\pi_i}$ en w_i. Ce calage "direct" réalise simultanément une correction de non-réponse (et génère donc une phase supplémentaire) et une réduction de la variance (calage sur des données externes). Si on considère que les probabilités de réponse sont estimées par $\hat{p}_i^{-1} = g_i = F\left(x_i' b\right) = \frac{w_i}{d_i}$ (voir F. DUPONT (1993)), on risque d'obtenir des probabilités supérieures à 1. Une étude particulière est nécessaire pour savoir si les formules développées précédemment s'appliquent encore. Dans le cas de données corrigées implicitement de la non-réponse par CALMAR, voici la démarche proposée dans la première version du logiciel POULPE :

- considérer que la probabilité de réponse est constante et égale à $p = \frac{r}{n}$ où r est le nombre de répondants et n la taille de l'échantillon.
- traiter le fichier avec une phase supplémentaire en considérant que la dernière phase est un sondage poissonnien de paramètre p et que les poids ont été modifiés par CALMAR.

CARON, N. : « Calcul de l'effet de sondage (Design Effect) dans le logiciel POULPE », note interne n°981/F410, 1996.

CARON, N. : « Estimations de variance négatives obtenues à partir de l'enquête Situations Défavorisées », notes internes n°976 et 983/F410, 1996.

DEVILLE, J.-C. : « Estimation de précision de données d'enquêtes », *document de travail Insee de la Direction des Statistiques Démographiques et Sociales* n°F9211, 1992.

DEVILLE, J.-C. : Support de cours TES (DAT 202-F) sur la non-réponse et le calage, 1997.

DEVILLE, J.-C. : « Estimation de la variance pour les enquêtes en deux phases », note interne manuscrite, Insee, 1993.

DEVILLE, J.-C, SÄRNDAL, C.-E. :« Calibration estimators in survey sampling », *JASA*, vol 87, n° 418, 1992.

DEVILLE, J.-C, SÄRNDAL, C.-E., SAUTORY, O. :« Generalized raking procedures in survey sampling », *JASA*, vol 88, n° 423, 1993.

DEVILLE, J.-C., VITE SAN-PEDRO C. : Rapport de recherche, Insee, 1993.

DUPONT, F. : « Calage et redressement de la non-réponse totale : validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989 », *Insee Méthodes,* n° 56-57-58, (Actes de journées de méthodologie Statistique de décembre 1993).

DUPONT, F. : « Éléments de spécifications pour la prise en compte de la non-réponse et des sondages en plusieurs phases dans le logiciel de calcul de précision des enquêtes effectuées par sondage », notes internes n°687/F010 et 4/F420, Insee, 1994.

DURBIN, J. : « Some results in sampling theory when the units are selected with unequal probabilities », *JRSS, serie B,* n°15, 1955.

RAJ, D. : « Some remarks on a simple procedure of sampling without replacement », *JASA*, n°61, 1966.

RAO, J.N.K. : « Unbiased variance estimation for multistage designs », *Sankhya,* C n°37, 1975.

ROSEN, B. : « Variance estimation for systematic pps-sampling », *rapport n°1991:15* de Statistique Suède, 1991

SÄRNDAL, C.-E., SWENSSON, B., WRETMAN, J. : *Model Assisted Survey Sampling*, Springer-Verlag, 1992.

LE LOGIEL POULPE : MODÉLISATION INFORMATIQUE

Jean-Noël Petit

Ce logiciel s'adresse aux responsables d'enquête soucieux d'évaluer la qualité des données recueillies sur des échantillons tirés de sondages complexes, à plusieurs degrés et plusieurs phases. Il apporte une estimation de l'erreur due à l'échantillonnage, en estimant la variance de l'estimateur d'Horvitz-Thompson sur les variables d'intérêt, et fournit l'intervalle de confiance à 95% centré sur le total pondéré.

Développé à partir du langage Sas Macro, il devrait être disponible prochainement sur les sites Mvs de l'Insee, et dans l'environnement Sas Windows, dans une version de pré-production.

1. Sur quelles bases reposent les calculs ?

Le logiciel s'appuie sur les informations issues de 3 sources différentes :

- le fichier des données résultant de l'enquête (appelé DATA), contrôlées et corrigées, éventuellement redressées à partir de données exogènes, par un logiciel approprié (par exemple Calmar),
- le fichier décrivant le plan de sondage (appelé MODELE),
- un fichier auxiliaire, ou fichier géographique (appelé GEO), avec les effectifs des unités administratives ou statistiques (appelées entités géographiques), destinés au calcul des probabilités d'inclusion.

Ces probabilités d'inclusion, indispensables au calcul des variances des estimateurs, figurent rarement dans le fichier de l'utilisateur, et doivent donc être évaluées par le logiciel, pour les différents degrés du sondage, à partir de :

- n : la taille de l'échantillon, présent implicitement dans le fichier de données,
- N : la taille de la population dans laquelle on a effectué le tirage : elle ne figure généralement pas dans le fichier de données, mais dans une source auxiliaire (GEO) ; cette source donne les effectifs de toutes les entités géographiques intervenant dans le tirage,

- le type de tirage : aléatoire simple, systématique, proportionnel à la taille, exhaustif,...

- des données auxiliaires, par exemple la taille pour les sondages proportionnels à la taille, ou les variables de tri pour les sondages systématiques...

Les formules utilisées permettent de s'affranchir des probabilités d'inclusion d'ordre 2, du type π_{ij}.

Pour ce calcul, le logiciel rapproche les sources d'information DATA et GEO à partir des identifiants des unités tirées, qui doivent donc être présents dans ces 2 sources. Il peut y avoir là un travail préparatoire à effectuer pour harmoniser les identifiants de ces sources, dans leurs dénominations et leurs types.

Pour les enquêtes ménages tirées de l'échantillon maître du RP 90, l'unité Méthodes Statistiques a constitué le fichier géographique de toutes les unités présentes dans ce sondage complexe : départements, communes, régions, strates unités primaires... avec leurs effectifs.

Une fois déterminées les probabilités d'inclusion, on effectue le calcul des variances de l'estimateur de Horvitz-Thompson, en s'appuyant sur un modèle de représentation (MODELE) apte à étendre à plusieurs degrés les formules connues pour les sondages à 1 degré.

2. Le modele sous-jacent au logiciel : comment represente-t-on un sondage a plusieurs degres ?

Lorsque l'on s'intéresse au calcul de la variance de l'estimateur de Horvitz-Thompson, on dispose d'un arsenal de formules pour un tirage à 1 degré, en fonction du type de tirage.

Exemple 1 : sondage à 1 degré proportionnel à la taille

Dans l'exemple ci-dessous, on tire des communes dans un canton proportionnellement à la taille de la commune.

On estime la somme d'une variable 'au niveau commune' y au niveau du canton par l'estimateur d'Horvitz-Thompson :

$$\hat{t} = \sum_s \frac{y_k}{\pi_k}$$

On connaît l'estimateur de la variance de cette somme, qui vaut :

$$\hat{V} = \frac{n}{n-1} \sum_k (1 - \pi_k)(\frac{y_k}{\pi_k} - \sum_s a_k \frac{y_k}{\pi_k})^2$$

avec : $a_k = \frac{(1-\pi_k)}{\sum_s (1-\pi_k)}$ (π_i : probabilité d'inclusion)

π_i = n* (Taille *de l'entité tirée) / Σ(Taille des entités tirables)*

soit : $\pi_i = n^*$ *(Effectif de la commune) / (Effectif du canton)*

On représente dans le logiciel un tel sondage par un arc sur lequel figurent ces informations (type de sondage, noms des entités), et on identifie les extrémités de l'arc par un code qui permettra de faire le lien entre les sondages élémentaires successifs.

On tire des communes dans des cantons proportionnellement à leur taille. On désigne par les termes :

- *entité tirée* : la commune (Enti),
- *entité d'agrégation* : le canton (Entag),
- *type de tirage* : tirage à probabilités proportionnelles à la taille (ppt).

Canton

AA

Entag

ppt

Enti

Commune

AB

On associe à chaque sondage un ensemble de données nécessaires au calcul des estimateurs :

- le type de tirage élémentaire : aléatoire simple, systématique,...
- les variables utilisées pour le tirage (par exemple la taille pour les tirages à probabilités proportionnelles à la taille),

- des données annexes relatives au sondage, et intervenant dans les formules : taux de sondage, population totale,... ; certaines de ces données sont fournies à part, d'autres (par exemple la population de l'échantillon) sont évaluées à partir de la base des données (DATA),
- entité tirée (ENTI) (canton, commune, groupe de communes, logement,...) ; c'est au niveau de cette entité que l'on applique les formules d'estimation liées aux éventuels sondages ultérieurs,
- entité d'agrégation (ENTAG) dans laquelle on tire les entités ENTI ; c'est à cette entité (ENTAG) que l'on applique les fonctions d'estimation,
- les variables d'intérêt sur lesquelles on applique les formules.

Exemple 2 : sondage à 2 degrés

Dans ce plan de sondage, on considère un arrondissement composé de 2 cantons :
- le canton 1, dans lequel on tire 3 communes : COM1, COM2, COM3
- le canton 2, dans lequel on tire 2 communes : COM4, COM5.

Puis, dans les communes du canton 1, on tire des logements par un tirage systématique (logements LOG1 à LOG7), dans celles du canton 2, des logements par un tirage aléatoire simple (logements LOG8 à LOG12).

Pour modéliser ce nouveau sondage, à 2 degrés, on étend le modèle précédent en y adjoignant 2 arcs, nommés AB-AC et AB-BC, pour obtenir finalement un «arbre» composé de :

- 3 arcs (AA-AB, AB-AC, AB-BC),
- 2 noeuds (AA et AB) et 2 feuilles (AC et BC).

Pour l'arc AA-AB, l'entité tirée est la commune, l'entité d'agrégation est le canton ; pour les arcs AB-AC et AB-BC, ce sont le logement et la commune.

CANT1 CANT2

Arbre des données

PPT PPT

COM 1 COM 2 COM 3 COM 4 COM 5

SYS SYS SYS SAS SAS

log1 log2 log3 log4 log 5 log 6 log 7 log 8 log 9 log 10 log 11 log 12

AA CANTON .....ya ....Va

Entag

PPT

Arbre des traitements

Enti

AB COMMUNE .....yb ....Vb

Entag

SYS SAS

Enti Enti

AC LOGEMENT BC LOGEMENT .....yc

On désigne par :

* y_c les valeurs de la variable d'intérêt y au niveau des feuilles AC et BC,

* y_b et V_b les variables " estimateur de total " et "estimateur de variance " au niveau du noeud AB,

* y_a et V_a les variables " estimateur de total " et "estimateur de variance " au niveau du nœud AA.

Lors du traitement des arcs AB-AC et AB-BC, on évalue les estimateurs yb et Vb en fonction des yc. On traite ensuite l'arc AA-AB en calculant les estimateurs y_a et V_a en fonction des y_b et des V_b.

Cette évaluation met en jeu des formules qui dépendent du type de tirage. Les y_b et V_b une fois estimés, deviennent les variables d'intérêt pour le sondage du niveau supérieur, et on renouvelle le processus jusqu'à aboutir à la racine de l'arbre (inversé).

Sur cet exemple, on obtient ainsi :

$$y_{com1} = f_{sys}(y_{lg1}, y_{lg2})$$

$$V_{com1} = g_{sys}(y_{lg1}, y_{lg2})$$

$$y_{com4} = f_{sas}(y_{lg8}, y_{lg9}, y_{lg10})$$

$$V_{com4} = g_{sas}(y_{lg8}, y_{lg9}, y_{lg10})$$

$$y_{cant1} = f_{ppt}(y_{com1}, y_{com2}, y_{com3})$$

$$V_{cant1} = g_{ppt}(y_{com1}, y_{com2}, y_{com3}) + g_{ppt}{}^{*}(V_{com1}, V_{com2}, V_{com3})$$

Les fonctions f_{sys} , g_{sys} ..., f_{ppt} , g_{ppt} et g^{*}_{ppt} qui représentent des formules de calcul appropriées à chaque type de tirage, sont la traduction de la formule de Raj :

$$\hat{V}(\hat{Y}) = f(\hat{t}) + \sum_{s} w_{is} \hat{V}_i$$ (se reporter à l'article de N. Caron)

De par les propriétés de l'estimateur de Horvitz-Thompson, elles s'expriment dans le logiciel par la formulation suivante :

la variance au niveau p est la somme de 2 termes :

- *la variance des sommes obtenues au niveau p-1,*
- *la somme pondérée des variances obtenues au niveau p-1, en utilisant les formules du sondage permettant de passer du niveau p au niveau p-1.* (p=1 pour les feuilles et p=n pour la racine de l'arbre).

Généralisation à n degrés

On peut étendre à n degrés le modèle retenu pour les sondages à 2 degrés, en reliant les divers arcs des sondages élémentaires : on aboutit ainsi à l'arbre modélisant le sondage à plusieurs degrés (voir en annexe 1 à titre d'exemple, une branche de l'arbre modélisant l'échantillon maître 1990 pour les communes de moins de 20 000 hab.)..

Le calcul s'exerce sur les données du fichier pour produire les premiers estimateurs (en général estimateur du total et de la variance pour n variables d'intérêt) relatifs aux derniers sondages élémentaires. Ces estimateurs deviennent les variables d'intérêt du sondage supérieur représenté aussi par un arc, avec des entités tirées qui sont les entités d'agrégation du sondage inférieur.

Le calcul récursif est conduit dans l'ordre chronologique inverse des opérations de tirage des entités : on commence par traiter les niveaux inférieurs (c'est-à-dire les dernières entités tirées) pour remonter ensuite aux niveaux supérieurs.

Une stratification est assimilée à un tirage avec un taux de sondage égal à 1.

3. Les différentes étapes

Après une phase préparatoire de mise en cohérence des sources, les traitements sont groupés en deux étapes.

Un premier ensemble d'étapes au cours desquelles :

- l'utilisateur décrit le plan de sondage au niveau de chaque arc, sous la forme d'une table SAS (appelée MODELE) ; cette modélisation du sondage demande une parfaite connaissance du plan de sondage ;

- le logiciel calcule les probabilités d'inclusion élémentaires (c'est-à-dire relatives à un sondage élémentaire), lorsqu'elles sont absentes du fichier de données, et globales (résultant des différents tirages successifs), à partir d'une table SAS

(appelée GEO) dans laquelle on aura inscrit au préalable les populations de toutes les unités tirées (ou leur taille pour les sondages à probabilités proportionnelles à la taille).

Cette étape est réalisée une seule fois.

Un deuxième ensemble d'étapes pour l'application des formules :

- on précise les statistiques d'intérêt : totaux de variables ou statistiques complexes,
- le logiciel calcule les variances estimées de ces statistiques.

Cette étape est relancée à chaque fois que l'on étudie de nouvelles statistiques. Elle fait appel à un ensemble de modules lancés individuellement ou générés à partir de noms d'étape globale : ESTIVAR (pour l'estimation des variables) ou ESTIFON (pour l'estimation de statistiques complexes).

On lance toutes les étapes à partir d'une interface qui permet de :

- passer les noms des fichiers,
- sélectionner les variables d'intérêt, les variables explicatives, les variables de calcul (poids, phases...),
- définir les paramètres d'exécution et sélectionner les traitements.

A) Préparation des fichiers

Il s'agit d'introduire dans les fichiers les identifiants qui permettront leur rapprochement.

Le fichier géographique (GEO)

Le fichier géographique apporte des informations auxiliaires comme les effectifs des unités sondées, ou leur taille, données nécessaires au calcul des probabilités d'inclusion. Ces données ne peuvent, en général, pas être générées à partir du fichier d'enquête, et proviennent donc de sources annexes (par exemple, le recensement de la population). Il faudra généralement harmoniser ses identifiants géographiques et ceux du fichier de données : mêmes types (numérique ou caractère sur n positions) et mêmes codes, comme l'exige le logiciel sous-jacent Sas.

Le fichier de données (DATA)

Pour pouvoir rapprocher le fichier de données du modèle représentant le sondage, il faut préciser à quel arc terminal de l'arbre des traitements se rapportent les données,

en mentionnant dans une variable supplémentaire (appelée *NINFFIC*) le code de la feuille concernée (code à 2 lettres). Ainsi, dans l'exemple de la page 5, le logement *log6* recevra le code feuille "AC" , le logement log11 recevra le code "AC".

On injecte ce code dans chaque observation de la table des données SAS, à partir des identifiants géographiques.

Pour les enquêtes en plusieurs phases, aux données recueillies sur le terrain, il faut adjoindre les données des échantillons des phases précédentes qui, bien que mises à zéro par le logiciel, sont indispensables dans les calculs.

L'utilisateur doit en outre fournir dans la base des données (DATA) :

- la probabilité d'inclusion de la 2ème phase, pour les enquêtes en 2 phases Poissonnien,
- la probabilité d'inclusion de la 3ème phase, pour les enquêtes en 3 phases,
- le poids final *Wip* de l'enquête, retenu pour la diffusion des résultats ; ce poids pourra (lorsqu'il y aura eu calage) différer du poids global élaboré par le logiciel à partir du plan de sondage.

C'est ce poids *Wip* qui est retenu pour l'évaluation des totaux « pondérés », sur lesquels sont centrés les intervalles de confiance produits par le logiciel.

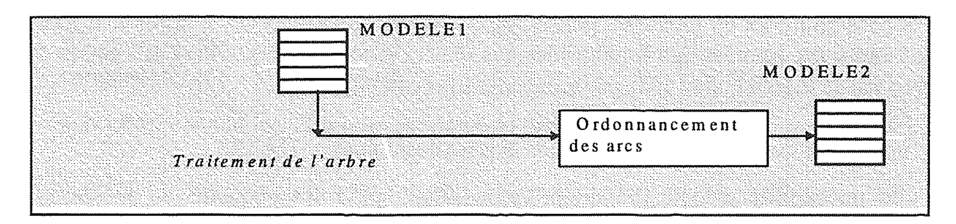
Le modèle du sondage (MODELE)

On crée la table Sas décrivant le modèle du sondage, chaque observation correspondant à un arc décrivant un tirage élémentaire, avec les données suivantes, au minimum :

- identifiants des extrémités de l'arc, sous la forme de codes à 2 lettres,
- type de tirage élémentaire : aléatoire simple, aléatoire simple équilibré, systématique, à probabilités d'inclusion inégales, exhaustif, total,
- noms des entités tirées : ex. REGION CANTON COMMUNE,
- noms des entités à l'intérieur desquelles on a réalisé le tirage : ex. REGION CANTON

B) Contrôle du modèle et génération

Une fois saisi le modèle, un module permet d'en vérifier certaines propriétés (connexité, absence de réseau...),

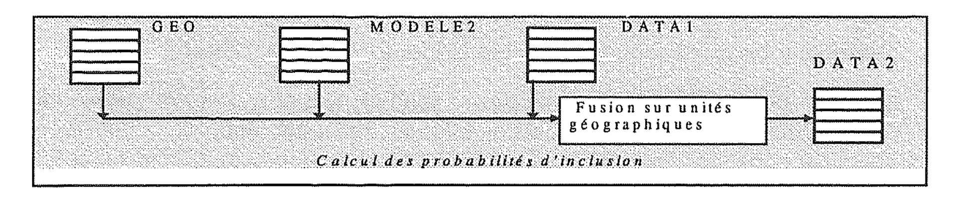


et de générer de nouvelles variables sur la topologie de l'arbre : la structure des formules impose de traiter les arcs dans un certain ordre pour suivre la règle suivante :

On peut traiter un arc lorsqu'il aboutit à une feuille ou lorsque tous ses arcs inférieurs ont été traités.

C) Calcul des probabilités d'inclusion

Dans cette étape, le logiciel calcule les probabilités d'inclusion élémentaires de la première phase, i.e. celles relatives à un sondage élémentaire, à partir :



- de données contenues dans le fichier de données : nombre d'entités tirées,
- de données issues du fichier "géographique" : taille des entités tirées, taille des entités d'agrégation,
- du type de sondage élémentaire, lorsque ces probabilités sont absentes du fichier de données.

Cas où les probabilités d'inclusion dépassent 1 : pour les sondages proportionnels à la taille, le calcul brut peut conduire à des valeurs supérieures à 1. Les données sont alors triées par ordre décroissant de taille des entités tirées ; on met à 1 la plus grande probabilité qui dépasse 1, et on recalcule les probabilités d'inclusion des autres entités tirées, après avoir mis à jour la taille de l'échantillon et la taille de l'ensemble des entités ; si à nouveau une valeur de probabilité dépasse 1, on réitère le processus.

Une fois toutes les probabilités d'inclusion élémentaires évaluées, leur produit fournit la probabilité globale d'inclusion de la première phase (égale à l'inverse du poids de sondage).

Pour les enquêtes en plusieurs phases dont la 2ème phase est stratifiée, le logiciel calcule la probabilité d'inclusion de la 2ème phase comme le rapport *nh/NH, nh* étant l'effectif pour la strate *h* dans l'échantillon 2ème phase, et *NH* l'effectif pour la strate *h* dans l'échantillon 1ère phase, ce qui requiert la présence des échantillons de la ou des phases précédentes dans la base des données de l'enquête.

Pour un sondage Poissonnien, le taux de sondage de la 2ème phase doit être présent dans le fichier de l'enquête (DATA).

La probabilité d'inclusion finale est égale au produit des probabilités d'inclusion relatives à chacune des phases.

D) Définition des variables d'intérêt

Les variables d'intérêt sont entrées sous la forme d'une liste de paramètres passée à une macro. La syntaxe du logiciel SAS est acceptée pour les variables de groupe, par exemple x1-x5 ou a--d ; cependant le nom d'une variable ne doit pas dépasser 7 caractères (pour une variable simple), ou 16 caractères (pour une variable de groupe) .

C'est dans cette étape que l'on définit les variables d'intérêt sur lesquelles on veut lancer le calcul des estimateurs, et que l'on prépare le fichier de données :

- en mettant à zéro les variables d'intérêt de la première phase (pour les enquêtes en 2 phases),

- en mettant à zéro les variables d'intérêt des 2 premières phases (pour les enquêtes en 3 phases).

E) Statistiques complexes : comment estimer des ratios ?

Les statistiques complexes (appelées fonctions dans ce document) sont traitées à l'aide d'un mécanisme particulier, qui permet de les décrire par appel de fonctions élémentaires (par exemple les fonctions arithmétiques somme, différence, produit, ratio ou exponentielles).

Base théorique de l'évaluation des estimateurs pour les fonctions

Etant donné un échantillon s d'une population P, il s'agit d'évaluer la précision d'une fonction construite sur des totaux : par exemple le revenu moyen par personne, à partir du revenu et du nombre de personnes du ménage.

Le logiciel procède par linéarisation des fonctions à estimer, à l'aide de la formule de Taylor, de développement en série à partir des dérivées partielles, approche formalisée par Woodruff en 1971, (démarche suivie par l'institut Statistics Sweden dans le logiciel CLAN). Il reprend également les principes de programmation du logiciel CLAN, à partir de fonctions élémentaires écrites en macro SAS et de fonctions "utilisateurs" bâties sur ces fonctions dérivables.

Ainsi, on remplace une fonction de totaux, par le total d'une fonction définie sur chaque observation, fonction linéaire des variables observées ; on sait alors estimer la variance de ces totaux de fonctions linéaires.

Cette démarche est pertinente pour les échantillons suffisamment représentatifs pour que l'on puisse négliger dans le développement en série de Taylor, les termes de rang supérieur ou égal à 2.

Soit, par exemple, à évaluer la fonction : $\theta = \frac{t1}{t2}$ (revenu moyen par ménage), avec :

* y1 représentant le revenu du ménage, t1 = Σy1,

* y2 représentant le nombre de ménages, t2 = Σy2.

On pose : $z_k = \frac{\partial \theta}{\partial t1} * y_{1k} + \frac{\partial \theta}{\partial t2} * y_{2k}$

Alors : $z_k = \frac{1}{\hat{t}_2} * y_{1k} - \frac{\hat{t}_1}{\hat{t}^2_2} * y_{2k}$

On démontre que la variance du total de z_k sur la population est approximativement égale à la variance de θ, c'est-à-dire à l'erreur quadratique moyenne, si l'on néglige le biais.

Ainsi on linéarise les fonctions par la formule de Taylor :

$$\hat{\theta} - \theta = \sum_{j=1}^{J} f_j^{'}(t)(\hat{t}_j - t_j) \text{ où } f_j'(t) = \frac{\partial f(t)}{\partial t_j}$$

et la formule de transformation de Woodruff :

$$z_k = \sum_{j=1}^{J} f_j'(\hat{t}) y_{jk}$$

puis, à l'aide de fonctions de base (+,-,*,/), on génère les calculs d'estimateurs à effectuer. Ainsi, pour toute fonction rationnelle que l'on peut écrire à partir des opérateurs de base, on peut exprimer la dérivée à partir :

- des règles de base pour la dérivation de fonctions de fonctions,

- de la dérivation de quatre fonctions arithmétiques de base (addition, soustraction, multiplication, division) ou d'autres fonctions (exponentielle).

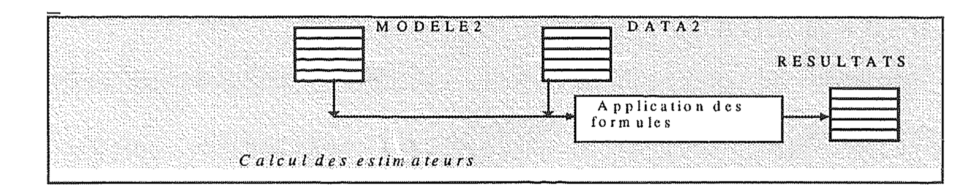
Définition d'une statistique complexe par l'utilisateur

Il reste à la charge de l'utilisateur la définition de la statistique complexe, à l'aide d'instructions qui se présentent ainsi :

%DIV(RATIO,REVENU,POP), si l'on cherche à mesurer le revenu moyen (RATIO), à partir de la variable REVENU et de l'indicatrice POP présentes dans le fichier : cette instruction permet de créer la variable z_k , dont la variance donnera celle de l'estimateur de la variable RATIO = REVENU/POP .

F) Calcul des estimateurs

Des modules spécifiques déroulent les formules de calcul pour obtenir :



- l'estimateur du total de la variable ou l'estimateur du total de la statistique complexe, à partir des probabilités d'inclusion du logiciel, qui ne tiennent pas compte d'éventuels calages,
- l'estimateur de la variance de cet estimateur,
- l'estimateur du total pondéré de la variable ou l'estimateur du total de la statistique complexe, à partir du poids final W_{ip},
- l'intervalle de confiance centré sur le total pondéré,
- l'effet de sondage ("design effect") sur option.

Ce calcul est fondé sur des formules propres au type de sondage, sur les variables d'intérêt présentes dans le fichier de données, sur les probabilités d'inclusion calculées en partie ou en totalité par le logiciel.

Le logiciel distingue 2 classes de formules :

- celles qui sont conduites au niveau des arcs du modèle, en suivant l'ordre défini par la règle du §2.b (ordre déterminé au cours de la génération du modèle), et qui font intervenir les probabilités d'inclusion élémentaires,
- celles qui sont appliquées sur l'ensemble des données, car elles ne font référence qu'aux probabilités d'inclusion globales.

Pour les enquêtes stratifiées, les formules exigent la création de variables intermédiaires au niveau de chacune des strates, pour chaque variable d'intérêt (ou statistique complexe), autant de variables 'strates' qu'il y a de strates, plus une.

Les valeurs manquantes sur les variables d'intérêt n'arrêtent pas les calculs, mais altèrent les résultats ; le logiciel en donne la fréquence.

On peut aussi faire appel à une méthode simplifiée de calcul des variances à partir des poids et d'un modèle dérivé du modèle théorique du plan de sondage.

Les formules de calcul aboutissent à des résultats erronés ou incomplets, dans les cas suivants:

- données manquantes,
- population de l'échantillon égale à 1.

Dans ces cas là, le logiciel met en oeuvre les traitements par défaut de Sas sur les données manquantes : elles sont ignorées.

4. Domaine d'application

Le logiciel couvre le domaine suivant :

- les sondages à 1 ou n degrés en une phase,
- les sondages à 1 ou n degrés, en deux phases, lorsque la seconde phase est un sondage aléatoire simple stratifié,
- les sondages à 1 ou n degrés, en deux phases, lorsque la seconde phase est un sondage Poissonnien,
- les sondages à 1 ou n degrés, en trois phases, lorsque la deuxième phase est un sondage aléatoire simple stratifié, et la troisième phase un sondage Poissonnien : ceci permet de traiter les enquêtes en deux phases stratifiées avec une correction de non réponse, en considérant cette dernière comme un sondage Poissonnien.

Pour les sondages élémentaires, les formules programmées actuellement portent sur les types suivants :

- sondage aléatoire simple sans remise,
- sondage à probabilités inégales (en particulier à probabilités proportionnelles à la taille),
- sondage systématique à probabilités égales,
- sondage stratifié,
- sondage équilibré.

La statistique d'intérêt peut être le total d'une variable, ou une statistique complexe, fonction de plusieurs totaux de variables. Dans ce dernier cas, la méthode de linéarisation, programmée dans le logiciel, permet de se ramener à l'estimation de la variance du total d'une variable synthétique.

Dans toutes ces configurations, le logiciel offre la possibilité de calculer les estimateurs d'Horvitz-Thompson et leur variance, et l'effet de sondage (Design Effect).

Pour les enquêtes dont les données ont été redressées par le logiciel Calmar, la précision est donnée par la variance des estimateurs sur les résidus, (et non plus sur les variables), calculés à partir d'une régression (procédure GML de SAS) sur les variables explicatives (numériques et caractères) passées au logiciel.

5. Mise en œuvre du logiciel

Les programmes sont écrits en langage Macro de SAS, avec appel de procédures courantes, et sans utilisation d'outil particulier de ce logiciel statistique en exécution. Ils tournent sous les systèmes d'exploitation MVS (IBM) et Windows (Microsoft).

La version actuelle est issue de spécifications d'études, et non de production. Elle dispose d'une interface interactive (en Sas Scl, ergonomie MVS), qui permet d'entrer les principaux paramètres, et de générer les macros d'exécution.

Principaux paramètres décrivant l'enquête

- nombre de phases de l'enquête : {1,2,3}, et nom de la variable phase,
- méthode de calcul : {simple, stratifié, Poissonnien, Ultimate Clusters},
- poids final,
- probabilités de réponse pour la 2ème phase Poissonnien, et la 3ème phase,
- liste de variables définissant les strates, pour les phases stratifiées,
- listes des variables explicatives pour les enquêtes ayant été redressées par le logiciel Calmar.

Les autres paramètres portent sur les noms des fichiers en entrée, la liste des variables d'intérêt et les options d'édition.

Ressources nécessaires

- espace disque : environ 6 fois la taille du fichier d'enquête (pour les fichiers intermédiaires, pour la création des fonctions, des variables strates, des résidus...),
- logiciel Sas version sur micro 6.11 sous Windows et 6.08 sur MVS.

Délais d'apprentissage et d'exécution du logiciel

L'opération de modélisation du sondage, lorsqu'il s'agit d'un sondage complexe, est plus longue que la mise en oeuvre du logiciel. En effet, le processus de modélisation, malgré son apparente simplicité, demande pour sa compréhension un minimum d'apprentissage.

Il faut compter environ une demi-journée pour se familiariser avec le lancement du logiciel ; en revanche les tâches préliminaires pourront demander davantage de temps, pour :

- déterminer avec rigueur les caractéristiques du sondage à l'origine de l'enquête,
- rassembler les sources contenant les effectifs nécessaires au calcul des probabilités d'inclusion,
- harmoniser les identifiants des diverses sources.

Pour les enquêtes de l'Institut issues du dernier échantillon-maître, l'essentiel de ce travail préparatoire a déjà été accompli, il ne reste généralement qu'à adapter à la marge le modèle, au niveau des derniers tirages.

Bibliographie

DEVILLE J.C. Estimation de précision de données d'enquête, Insee F9211, 1992 ;

Calcul de l'effet de sondage, Mars 96.

ISNARD M. Validation expérimentale du modèle théorique, Décembre 1992.

NEROS B. Base de sondage des logements neufs, 589/F010, Octobre 1993.

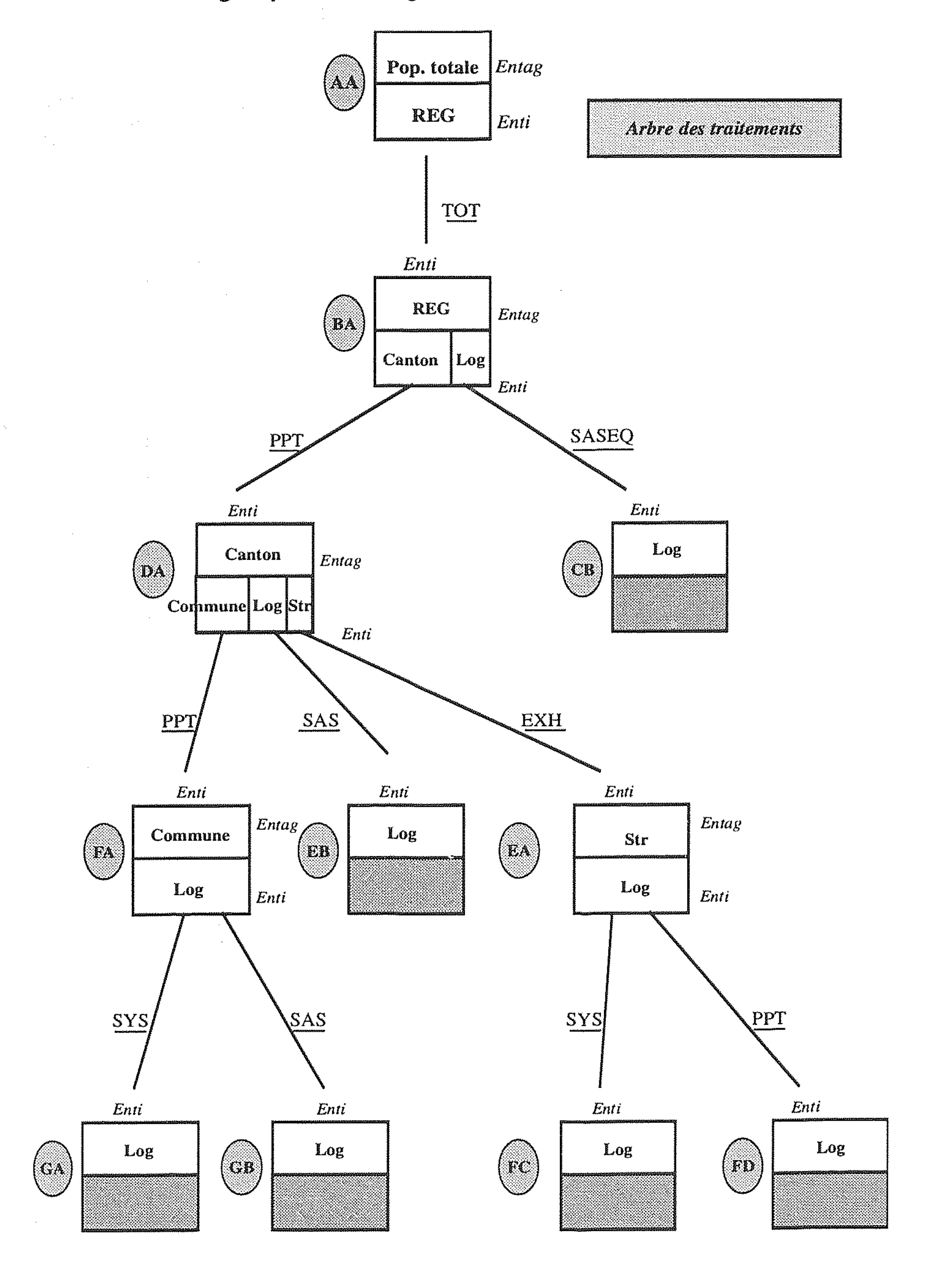
Projet sur l'estimation de précision :
fichier géographique, et fichier de données issu de l'enquête, 688/F010, Mars 1994.

SAUTORY O. Estimation de la variance dans un plan de sondage à plusieurs degrés, Insee, 3 Nov 92.

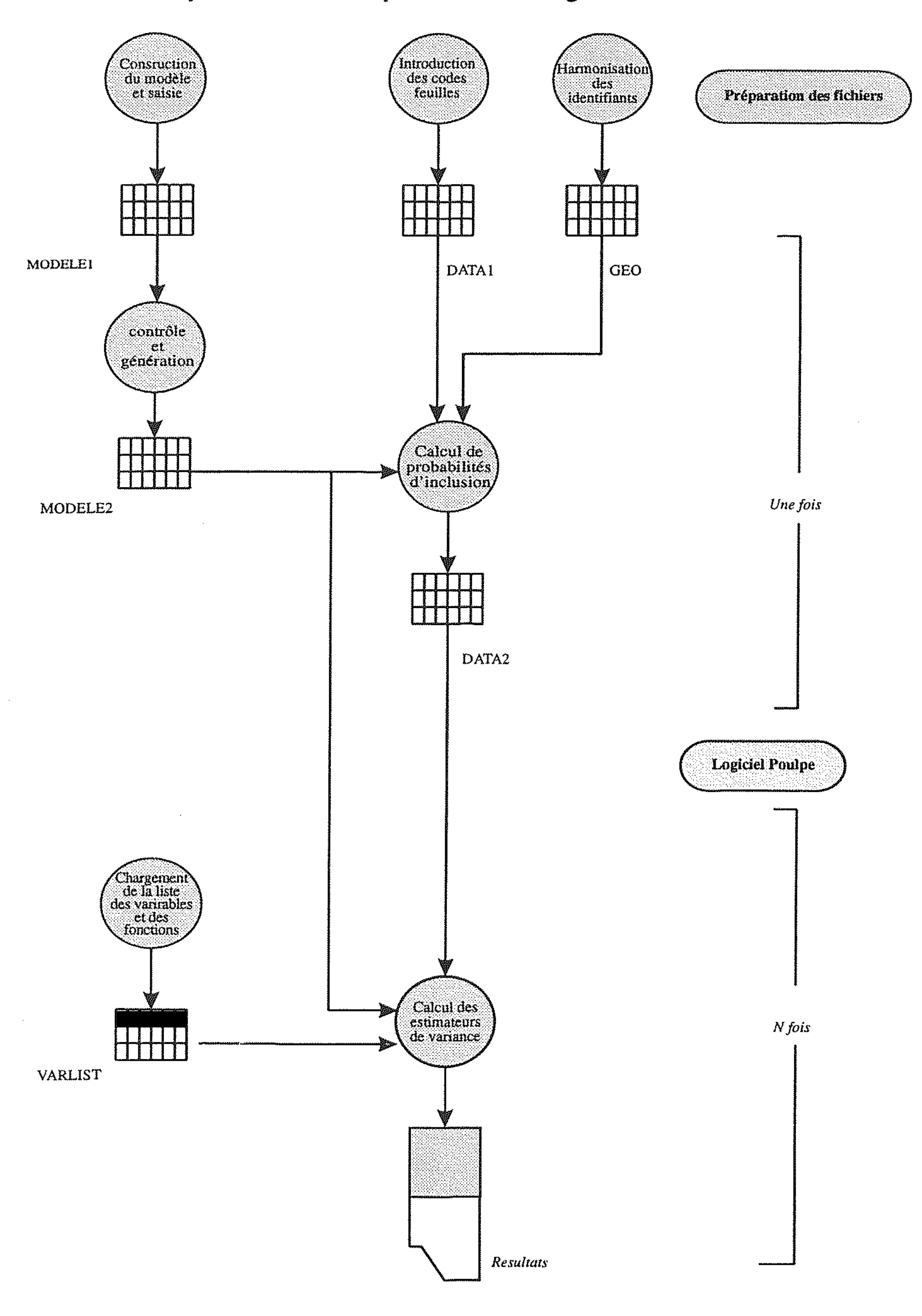
DUPONT F. Non réponse et sondages en plusieurs phases dans le logiciel de calcul de précision, Insee 4/F420, 10 Juin 94.

CARON N. Sondage 2ème phase Poissonnien dans le logiciel Poulpe
Insee 958/F410, 29 Jan 96.
Calcul simplifié de la variance, Insee 968/F410, 22 Février 96.

Modèle de sondage à plusieurs degrés



Estimation de précision des enquêtes : schéma général



UTILISATION DU LOGICIEL POULPE POUR LE CALCUL DE LA PRÉCISION D'ESTIMATEURS TIRÉS DE L'ENQUÊTE LOGEMENT 1996

David le Blanc

Introduction

Ce court article n'a pas de prétention méthodologique ; son ambition se borne à présenter l'application du logiciel POULPE à une enquête auprès des ménages.

Dans une première partie, on rappelle les particularités de l'enquête Logement par rapport aux autres enquêtes Ménages de l'Insee. La principale tient aux quantités que l'on cherche à estimer à partir de l'enquête : pour résumer, l'enquête Logement doit servir à estimer non seulement des structures, mais aussi des niveaux.

Le premier but de l'enquête Logement est de fournir un nombre de résidences principales, ou de ménages, mais aussi un nombre total de logements, les plus précis possibles. L'enquête sert également à donner des estimations du niveau de certaines variables portant sur le champ des ménages (ou des résidences principales) : nombre de propriétaires, de locataires privés et HLM, etc. Là encore, les niveaux ont de l'importance, car ces quantités doivent pouvoir être reliées à des grandeurs physiques, comme les flux d'aides à la pierre ou à la personne. Enfin, on calcule des estimateurs de type ratio, comme la part des propriétaires dans les ménages, de façon analogue à ce qui se pratique dans les autres enquêtes auprès des ménages.

La mise en oeuvre de POULPE suppose d'abord que l'on modélise le plan de sondage de l'enquête, afin de reconstituer les probabilités d'inclusion a priori de chacun des logements tirés. On doit ensuite modéliser la procédure de redressement et de calage. Vu la complexité du redressement de l'enquête, des simplifications sont nécessaires pour faire fonctionner POULPE.

Dans une deuxième partie, on s'intéresse aux résultats obtenus, selon deux optiques différentes :

- une perspective d'utilisation par les concepteurs de l'enquête. Quelle est la précision obtenue sur différents types d'indicateurs ? Comment cette précision

permet-elle d'interpréter les séries des enquêtes Logement depuis 12 ans ? Cette précision est-elle améliorée par le calage, et sur quelles variables ? Y a-t-il un « bon poids » ?

- une perspective plus méthodologique d'expertise de l'échantillon-maître et de la précision des différents types de variables dans les enquêtes tirées de cet échantillon. Il s'agit de calculer les précisions d'un grand nombre de variables, afin de pouvoir constituer des catégories, selon d'une part la perte de précision due au tirage dans l'EM (« design effect ») et d'autre part le gain de précision apporté par le calage sur marges (effet « CALMAR »). La comparaison de la précision des estimateurs calculée par POULPE avec celle qui résulte de calculs approchés permet en outre de donner des règles approximatives pour calculer la précision de variables quelconques, sans que le passage de POULPE soit nécessaire.

I) Plan de sondage et redressement de l'enquête Logement 1996

I-1) Plan de sondage

L'enquête Logement ne se distingue pas des enquêtes ménages traditionnelles au point de vue du plan de sondage. Deux bases de sondage sont utilisées : l'échantillon-maître (EM) et la base de sondage des logements neufs (BSLN). Le tirage s'effectue en une seule phase. Les taux de sondage sont les suivants :

Echantillon	Taux de sondage
EM :	
résidences principales et logements vacants non ruraux	1/ 720
logements secondaires ou occasionnels	1/ 1440
logements vacants en zone rurale	1/ 1080
BSLN (tous statuts):	1/ 361

I-2) Redressement

En revanche, en ce qui concerne le redressement, l'enquête Logement se distingue fortement des autres enquêtes ménages. Celles-ci sont généralement redressées en une seule étape, par calage de la structure de certaines variables de l'enquête sur les

marges de l'enquête Emploi la plus proche dans le temps. Ce calage est effectué sur des variables socio-démographiques comme le nombre de personnes, le nombre d'actifs, la strate géographique, la catégorie socioprofessionnelle et l'âge de la personne de référence du ménage. On considère en effet que l'enquête Emploi, avec un échantillon très important (plus de 100 000 logements), est la plus précise disponible à l'Insee ayant lieu à des intervalles de temps rapprochés.

Pour l'enquête Logement, on ne recourt pas à cette méthode, pour plusieurs raisons :

- un des résultats les plus attendus de l'enquête est l'évaluation du parc de logements et de ses différentes composantes (voir la **figure 1**). Il s'agit d'estimer un *nombre* de résidences principales, ou de ménages, mais aussi un nombre total de logements, les plus précis possibles. Ces quantités sont notamment utilisées dans une optique de suivi annuel du parc de logements et de ses différentes composantes, dans le compte satellite du Logement. Le fait qu'on cherche à relier les stocks de logements dans chaque catégorie aux flux annuels qui affectent ces catégories rend cruciale une bonne estimation du niveau des stocks, dans la mesure où les statistiques les plus fiables, celles du recensement de la population, sont très espacées dans le temps.

- le but premier des enquêtes Emploi n'est pas d'estimer précisément le parc de logements[1]. De fait, des divergences importantes existent entre niveau du parc de logements décrit par les enquêtes Emploi et celui donné par les enquêtes Logement. De plus, l'importance de l'échantillon de l'enquête Logement (40 000 logements) permet des raffinements en matière de redressement, de sorte que l'on peut penser que cette enquête estime le parc de logements aussi bien, sinon mieux, que l'enquête Emploi.

Des procédures de redressement spécifiques sont donc mises en oeuvre.

A) Correction de la non-réponse

La non-réponse est corrigée à l'aide du logiciel CALMAR. En se plaçant sur le champ des résidences principales, il s'agit d'abord de déterminer les variables (disponibles dans la base de sondage) les plus discriminantes du point de vue de la non-réponse. Les poids des observations répondantes sont ensuite modifiés de manière à respecter les marges de ces variables évaluées sur l'ensemble des résidences principales.

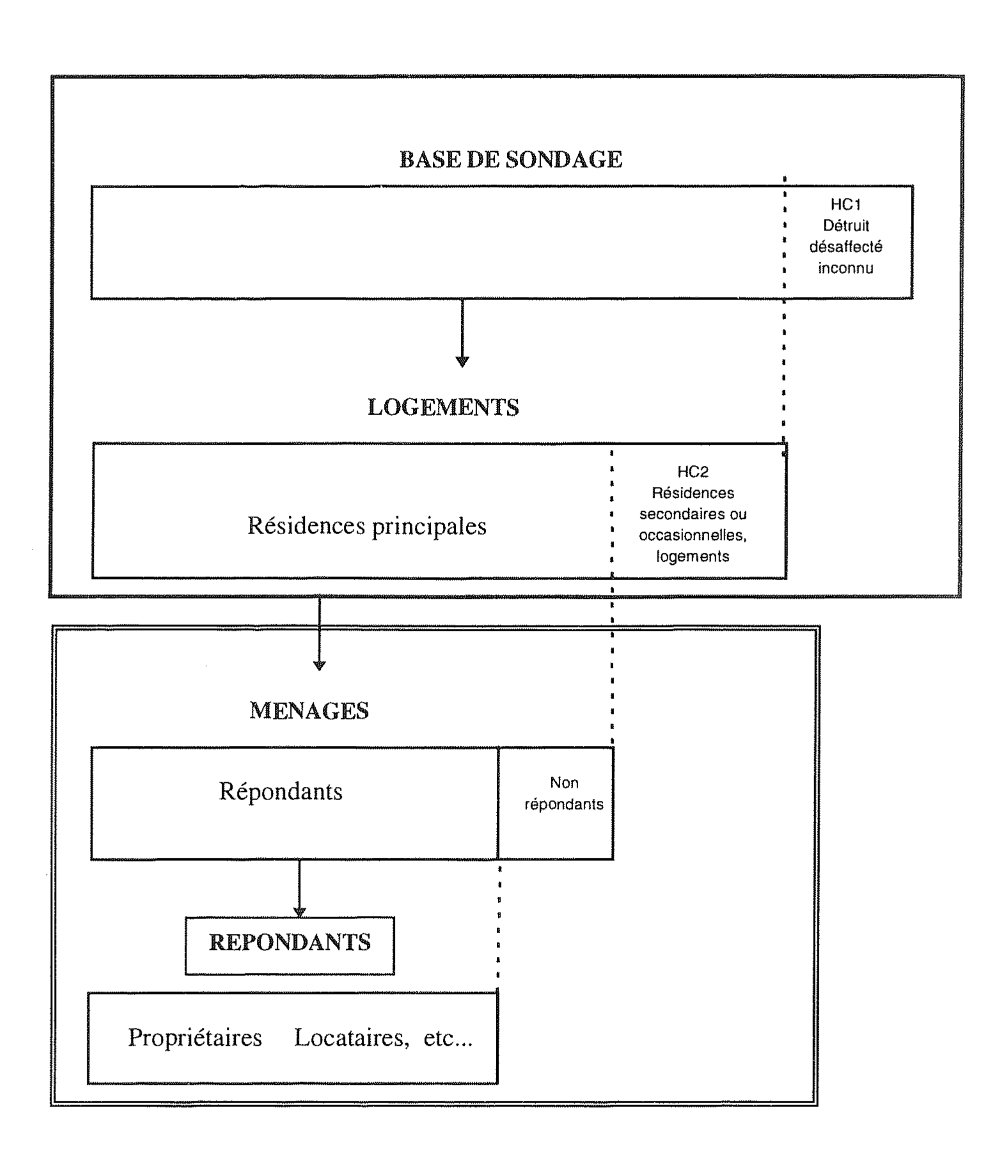
---

1 L'enquête Emploi elle-même n'est pas calée sur un nombre exogène de ménages, mais sur la population estimée, stratifiée par sexe et âge.

Figure 1
Les différences entre l'enquête logement et une enquête ménage traditionnelle

Encadré noir : étape d'évaluation du parc de logements (spécificité de l'enquête Logement)

Encadré double : étape d'évaluation de variables ménages (toutes les enquêtes ménages)



Compte tenu des informations différentes disponibles dans les bases de sondage, on distingue trois sous-échantillons :

- les logements principaux,
- les logements non principaux au recensement de 1990,
- les logements neufs.

Sur chaque sous-échantillon, les variables les plus discriminantes pour la non-réponse sont mises en évidence à l'aide d'un modèle LOGIT (voir la **figure 2**). La macro CALMAR sert ensuite à modifier les poids des observations répondantes.

B) Calage sur les marges : correction des aléas d'échantillonnage

La deuxième étape consiste à caler l'échantillon (complet cette fois, c'est-à-dire l'ensemble des logements quelle que soit leur utilisation) sur des marges externes afin de limiter au maximum les fluctuations d'échantillonnage. Les logements tirés de l'échantillon-maître sont calés à partir de variables enregistrées au recensement de la population, les logements issus de la BSLN sur des marges tirées du fichier SICLONE des permis de construire.

Les variables qui servent au calage ne sont pas les mêmes selon que le logement est tiré d'une des deux bases de sondages, et pour l'échantillon-maître selon que son statut au recensement était résidence principale ou non. Ces variables sont indiquées dans la **figure 2**.

C) Correction des effectifs des bases de sondage

Cette étape est destinée à corriger la non-exhaustivité des bases de sondage. D'une part, il faut tenir compte des réaffectations de locaux en logements survenues depuis le dernier recensement. D'autre part, afin de tenir compte des données les plus récentes sur la construction neuve (enregistrées à partir des permis de construire), les poids des logements neufs sont modifiés.

La procédure de redressement suivie peut sembler inutilement lourde ; les résultats donnés par POULPE permettent de quantifier son efficacité, c'est-à-dire le gain de précision qu'elle apporte. Comme on va le voir, la précision obtenue sur des variables-clés (nombre de logements et de ménages, nombre de propriétaires) justifie tout-à-fait cette procédure de redressement par rapport à l'alternative qui consisterait à se caler sur des marges de l'enquête Emploi.

I-3) Application du logiciel POULPE à l'enquête Logement

La complexité du redressement effectué nécessite une simplification pour la mise en oeuvre de POULPE : le plan de sondage est approximé en particulier pour la base de sondage des logements neufs, ainsi que les procédures de correction de la non-réponse et de calage sur marges (voir la **figure 2**).

A) Deux niveaux de calculs de précision

Les particularités présentées au I-2) font que deux types d'estimation de la précision sont nécessaires, correspondant aux deux niveaux de variables produites par l'enquête :

1) précision sur le nombre de logements des différentes catégories (résidences principales, secondaires ou occasionnelles, logements vacants). La base de sondage étant supposée exhaustive, les hors-champ sont uniquement constitués par les logements inconnus ou ayant disparu. Ce cas ne sera pas détaillé ;

2) précision sur des variables de type ménage. Ces variables sont estimées sur le champ des résidences principales uniquement. On est donc dans un cas d'application de POULPE similaire à ce que l'on peut rencontrer dans les autres enquêtes ménages. Rappelons que l'idée sous-jacente du calcul de la précision de ces variables est de se ramener à des variables artificielles dont l'estimateur de Horvitz et Thompson corresponde à l'estimateur utilisé après redressement.

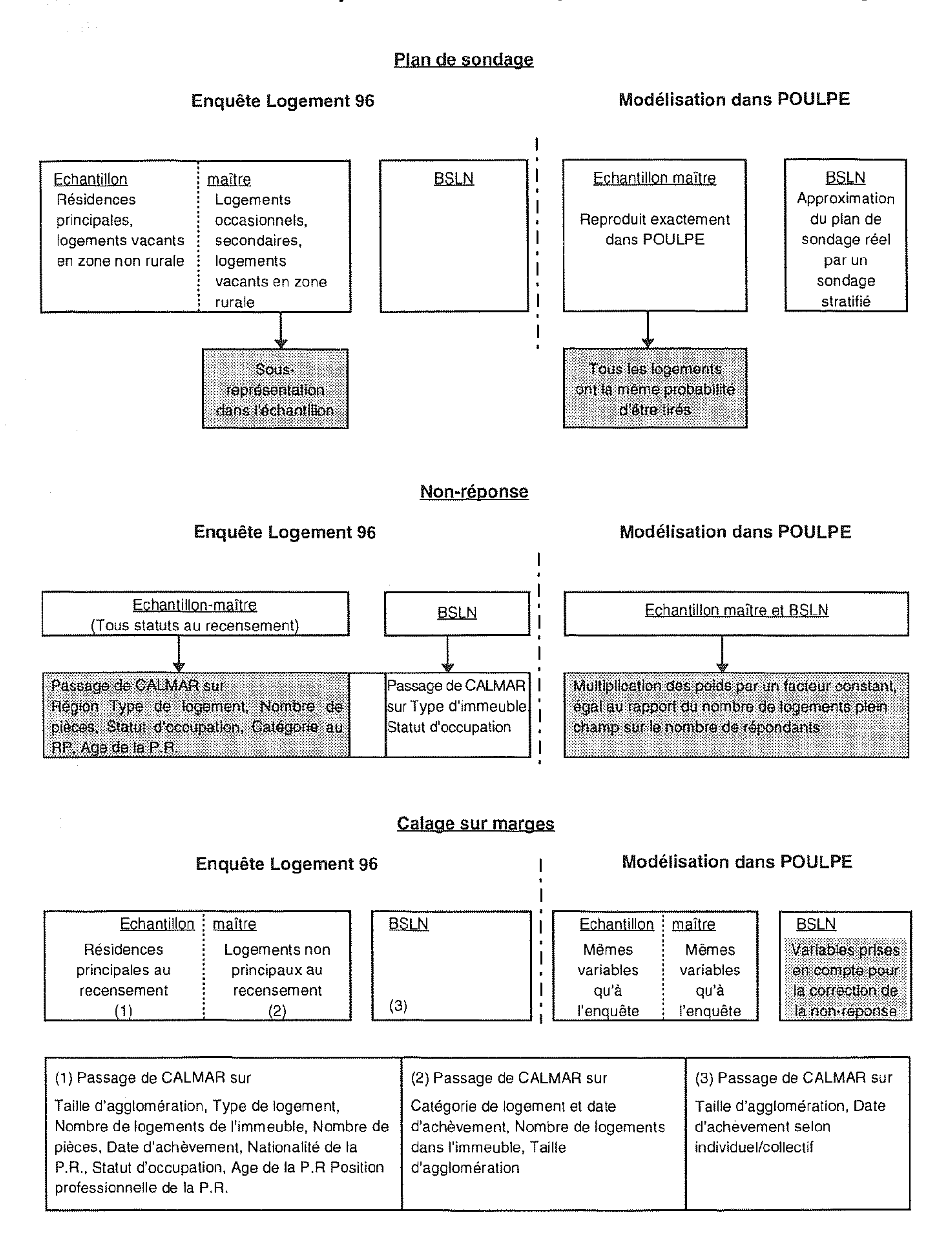
B) Précision sur des variables de type ménage

En ce qui concerne le **plan de sondage**, on considère que les logements issus de l'échantillon-maître ont tous la même probabilité d'être tirés, ce qui revient à négliger la sous-représentation des logements non-principaux. Pour les logements neufs, le plan complexe de la BSLN est approximé par un sondage stratifié, avec 33 strates (pour plus de précision, on se reportera au document méthodologique à paraître).

En ce qui concerne **le redressement**, deux approximations sont faites.

1) On considère que la non-réponse a été corrigée de façon globale, en multipliant les poids des observations répondantes par un facteur constant égal au rapport du nombre de résidences dans le champ de l'enquête sur le nombre de répondants. Pour pouvoir faire des calculs tenant compte de la correction de la non-réponse, on considère que cette correction constitue une phase supplémentaire dans le tirage de l'échantillon : les probabilités d'inclusion seront calculées comme si à partir de l'échantillon de logements, on avait tiré des résidences principales. Le logiciel POULPE sera donc paramétré comme pour un tirage en deux phases.

Figure 2
Plan de sondage et redressement de l'enquête Logement
Modélisation dans POULPE pour l'estimation de précision de variables ménages



(1) Passage de CALMAR sur	(2) Passage de CALMAR sur	(3) Passage de CALMAR sur
Taille d'agglomération, Type de logement, Nombre de logements de l'immeuble, Nombre de pièces, Date d'achèvement, Nationalité de la P.R., Statut d'occupation, Age de la P.R Position professionnelle de la P.R.	Catégorie de logement et date d'achèvement, Nombre de logements dans l'immeuble, Taille d'agglomération	Taille d'agglomération, Date d'achèvement selon individuel/collectif

A partir de ces données d'une part, et d'une modélisation de l'échantillon-maître et de la BSLN d'autre part, POULPE reconstitue le plan de sondage de l'enquête, en calculant les probabilités d'inclusion de tous les logements. Les probabilités a priori d'inclusion simple sont estimées en modélisant le plan de sondage de l'échantillon-maître par un arbre[2]. (se reporter à l'article de J.N. Petit).

Cette approximation de la non-réponse est évidemment grossière, et conduit sans doute à surestimer le gain de précision dû à la correction des biais d'échantillonnage.

2) On considère que le calage sur marge a été effectué indépendamment sur les trois sous-populations suivantes :

- logements principaux au recensement de 1990,
- logements non principaux au recensement de 1990,
- logements issus de la BSLN.

Pour les deux premières sous-populations, les variables utilisées dans POULPE sont celles qui interviennent dans le calage effectué pour corriger les aléas d'échantillonnage. Pour les logements neufs, les variables utilisées sont celles qui sont utilisées pour la correction de la non-réponse.

C) Résultats obtenus en sortie de POULPE

Pour chaque variable d'intérêt, trois calculs de précision sont réalisés :

① la précision obtenue sur les données brutes corrigées par un facteur constant correspondant au rapport du nombre de résidences dans le champ de l'enquête sur le nombre de répondants (simulation de correction de la non-réponse) et par conséquent avant le passage de CALMAR,

② la précision obtenue sur les données corrigées de la non-réponse et des fluctuations d'échantillonnage, c'est-à-dire après le passage de CALMAR ,

③ la précision obtenue sur les données corrigées de la non-réponse et des fluctuations d'échantillonnage en considérant que le plan de sondage est celui d'un sondage aléatoire simple (SAS). Ce calcul permet de comparer la variance obtenue par le plan de sondage complexe de l'échantillon-maître à celle que l'on aurait

---

2 Il faut noter que les probabilités d'inclusion simple sont calculées lors du tirage de l'échantillon des enquêtes ménages ; une amélioration évidente à apporter au calcul de précision des enquêtes consisterait à garder cette information pour alimenter le logiciel POULPE. On éviterait ainsi un calcul approximatif a posteriori. En pratique cependant, on peut vérifier que ces calculs donnent des résultats satisfaisants : les probabilités d'inclusion simple calculées par POULPE sont distribuées autour de leur valeur théorique.

obtenue si le plan de sondage avait été celui d'un SAS. Le rapport des deux variances estimées est appelé « design effect ».

Un moyen commode pour comparer le degré de précision de différentes variables consiste à comparer leurs coefficients de variation, sans unité. Pour évaluer le gain de précision dû à la correction de la non-réponse et au calage sur marges, on peut privilégier le rapport des variances avant et après le passage de CALMAR. Ce rapport peut en effet être interprété comme le gain réalisé en terme de taille d'échantillon, et donc de coût de l'enquête, pour un niveau de précision donné.

II) Résultats

Lors de l'examen des résultats, il importe de garder à l'esprit que les écarts-types estimés après calage sur marge calculés par POULPE sont basés sur l'hypothèse que les marges de calage sont connues sans erreur. Dans la pratique, cette hypothèse est couramment admise pour les chiffres tirés du recensement de la population ; pour ce qui est des logements neufs, cela est moins évident, dans la mesure où les statistiques tirées du fichier SICLONE du ministère de l'Equipement subissent des révisions importantes pendant plusieurs années après la date d'enquête.

II-1) Une optique « concepteur d'enquête »

A) Nombre de logements des différentes catégories

Les résultats numériques figurent dans le **tableau 1**. Ils confirment des idées déjà connues, mais que l'on ne pouvait quantifier auparavant.

- Le calage effectué lors du redressement a un effet important sur la précision de la mesure du parc de logements. L'écart-type sur le nombre de résidences principales est divisé par deux lors du redressement, ce qui signifie que la précision après redressement est celle d'un échantillon non redressé quatre fois plus grand.

- Le plan de sondage de l'échantillon-maître est très bon pour estimer les résidences principales, champ ordinaire des enquêtes ménages. Le design effect estimé après passage de CALMAR pour cette variable, d'une valeur de 1,23 , indique que la perte de précision par rapport à un sondage aléatoire simple est minime. Le nombre de ménages est connu (à 95 % de confiance) à plus ou moins 0,4 % près, ce qui représente environ 95 000 ménages. A titre de comparaison, la précision donnée par les concepteurs de l'enquête Emploi est de plus ou moins 105 000 ménages, pour un échantillon nettement plus important. La procédure de redressement sophistiquée est donc justifiée.

Tableau 1
Estimations de la précision par POULPE pour l'enquête Logement 1996.
Nombre de logements selon le statut

	Estimateur pondéré (milliers)	Ecart-type avant calage	Ecart-type après calage	Coefficient de variation après calage (%)	Design effect après calage
Nombre de résidences principales	23 286	106 893	48 208	0,21	1,23
Nombre de résidences secondaires	2 452	104 218	41 935	1,71	2,68
Nombre de résidences occasionnelles	252	14 998	14 633	5,81	1,34
Nombre de logements vacants	2 231	41 444	39 075	1,75	1,23
Nombre de résidences secondaires et occasionnelles	2 704	105 173	42 652	1,58	2,31
Total des logements	**28 221**	**53 864**	**20 155**	**0,07**	**1,07**

- L'échantillon-maître se révèle nettement moins performant pour estimer le nombre de résidences secondaires. Le plan de sondage de l'échantillon-maître n'est pas optimal pour estimer le nombre de ces résidences, qui sont très concentrées géographiquement et conduisent à un fort effet de grappe (et donc à une variance importante). Toutefois, le passage de CALMAR permet de diviser par 2,5 l'écart-type sur cette estimation.

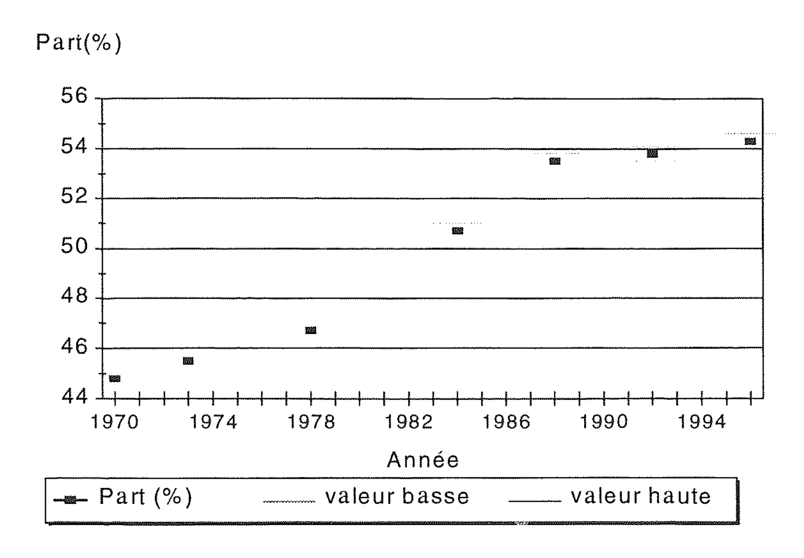
B) Variables ménages : parts des grands statuts d'occupation, flux quadriennaux

Indépendamment de la valeur intrinsèque d'estimations de précision pour une enquête donnée, disposer d'une estimation de la précision des données est nécessaire pour une exploitation correcte de la série des enquêtes Logement, particulièrement pour deux types de variables :

- la part des ménages dans les différents statuts d'occupation : propriété, location HLM, location libre. Un des résultats majeurs de l'enquête de 1992 avait été de mettre clairement en évidence que la part des ménages propriétaires, après une décennie d'augmentation rapide, demeurait stable, aux alentours de 54 %. Compte tenu de l'importance accordée à l'accession à la propriété dans les politiques du logement en France, ce résultat avait provoqué un certain émoi. Grâce à POULPE,

on sait que le passage de 53,8 à 54,3 entre 1992 et 1996 traduit plutôt une stabilisation qu'une reprise[3] (voir **graphique 1**) ;

Graphique 1
Evolution de la part des ménages propriétaires de leur résidence principale



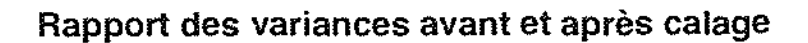
- les flux quadriennaux. Il s'agit de quantités mesurées à l'enquête 1996, portant sur des événements intervenues depuis la date de l'enquête précédente (novembre 1992). Ces flux permettent de mettre en cohérence les flux et les stocks entre les quatre dernières enquêtes. Dans la perspective d'utiliser la série des enquêtes Logement, qui ont lieu tous les quatre ans, pour suivre des générations d'une enquête à l'autre, on souhaite de même connaître la précision sur les tranches d'âge quadriennales.

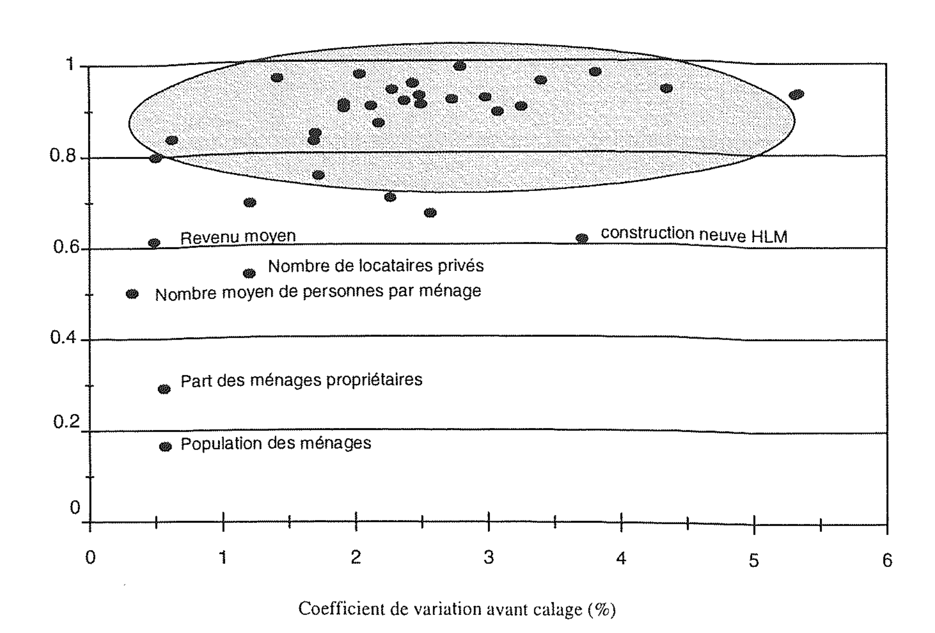
---

3. La mise à disposition du logiciel POULPE a provoqué un regain d'intérêt pour les estimations de précision à l'enquête Logement. Celles-ci étaient avant 1992 tombées en désuétude, alors même que dans le passé des estimations avaient été produites (pour l'enquête de 1973 notamment). Il est frappant de rapprocher le besoin de précision de l'enquête et l'évolution de la part des ménages propriétaires, chiffre certainement le plus « sensible » de l'enquête. En 1973, la hausse de cet indicateur par rapport à 1970 était très faible (+ 0,7 point) : cette hausse traduisait-elle une augmentation réelle ? Par la suite, la progression rapide de la propriété (+ 0,54 point par an en moyenne) faisait apparaître entre deux enquêtes successives une différence évidemment significative, qui ne nécessitait pas de calculs de précision. Ce n'est qu'en 1992, au vu d'une stagnation sur les quatre dernières années, que la question s'est de nouveau posée avec acuité. Entre-temps, les « calculs approché » avaient sans doute perdu des adeptes.

Une des questions sous-jacentes est celle de la nécessité d'opérer des redressements particuliers pour réaliser de telles exploitations : en effet, les données de flux ne font pas partie des variables calées, car on ne dispose pas de marges externes pour ces quantités. En revanche, il serait tout-à-fait envisageable de caler les effectifs des ménages par tranche d'âge sur des estimations de population exogènes.

Graphique 2
Amélioration apportée par le calage sur marges





Les estimations de précision concernant ces variables sont données dans **le tableau 2**. Si l'on met à part le cas de la construction neuve HLM pour laquelle l'estimation donnée par POULPE n'est pas bonne (voir plus bas), les trois autres variables de flux ont des « design effects » proches de 1 ; le passage de CALMAR n'apporte qu'un faible gain de précision. Le cas des variables indicatrices des tranches d'âge est significatif : la tranche des ménages de plus de 65 ans, proche des modalités de la variable de calage (plus de 75 ans), voit son écart-type divisé par deux après le passage de CALMAR. En revanche, les tranches d'âge 20-24 ans et 24-28 ans étaient incluses à l'intérieur d'une même modalité de la variable de calage (moins de 30 ans) : leur précision n'est pratiquement pas améliorée par le passage de CALMAR. Finalement, si l'on veut travailler sur des tranches d'âge quadriennales, il faudrait sans doute caler sur des marges respectant ces tranches.

Tableau 2
Estimation par POULPE de la précision pour des variables de type flux et des tranches d'âge quadriennales.

Variable	Valeur (milliers)	Design effect CALMAR	Coefficient de variation CALMAR	ρ*
Nombre de nouveaux ménages	2 185	0.99	1.57	0.85
Construction neuve HLM	268	0.56	2.91	0.61
Nombre d'accédants récents	1 658	0.98	1.82	0.90
Nombre de ménages ayant changé de logement dans la même commune depuis 1992	2 570	1.07	1.54	0.83
Nombre de ménages de plus de 65 ans	5 642	1.06	0.64	0.32
Nombre de ménages entre 20 et 23 ans	640	1.00	3.09	0.90
Nombre de ménages entre 24 et 27 ans	1 264	1.09	2.17	0.90

* ρ : rapport des variances après calage et avant calage. C'est une mesure de l'amélioration apportée par le calage.

C) Conclusion

Le redressement compliqué ne sert « qu'à » améliorer la précision des variables vitales pour l'enquête (nombre de logements et de résidences principales, part des propriétaires, etc.). Il faut noter que ce constat justifie dans une large mesure la pratique courante de calage des enquêtes ménages (du moins celles dont le logement n'est pas l'objet principal) sur un nombre de ménages exogène. L'effet du redressement sur d'autres variables est beaucoup plus limité, en particulier sur les flux quadriennaux qui constituent une originalité de l'enquête Logement.

II-2) Une analyse plus globale

Le choix des variables pour lesquelles on a utilisé POULPE n'est pas innocent. Dans une perspective d'analyse des résultats, il s'agissait de représenter différents types de variables : totaux de variables indicatrices et ratios correspondants (par exemple, nombre de ménages locataires et part des ménages locataires), variables de type financier (moyennes de revenu ou de loyer), ratios portant sur un sous-champ (par exemple, part des locataires évoluant en secteur HLM), modalités très peu fréquentes dans la population étudiée.

Il s'agissait aussi de constituer des groupes de variables selon leur précision, avant et après calage, l'idée étant, pour chaque type de variables ainsi déterminé, de

rechercher des règles approchées permettant le calcul de la précision d'une variable quelconque, POULPE étant encore coûteux en temps et en espace disque pour le moment.

De cette analyse, on peut tirer deux enseignements principaux.

1) La variance estimée par POULPE pour des totaux de variables indicatrices et les ratios correspondants, qu'ils portent sur l'ensemble des ménages ou sur un sous-champ, peut être très correctement remplacée par une estimation approchée ne prenant pas en compte le plan de sondage.

On s'intéresse à une caractéristique des ménages. Soit X la variable indicatrice relative à cette caractéristique, et **X** son total dans la population des ménages. Soit p la proportion de ménages ayant la caractéristique étudiée. Le statisticien estime **X** et p par les estimateurs pondérés :

$$\hat{X} = \sum_r w_i X_i \quad , \qquad \hat{p} = \frac{\sum_r w_i X_i}{\sum_r w_i}$$

où r désigne l'échantillon des répondants et $(w_i, i \in r)$ est le jeu des poids des observations répondantes.

Supposons maintenant que l'on soit dans le cas d'un sondage aléatoire simple à probabilités égales dans un échantillon r de taille n tiré d'une population de taille connue N. On suppose que toutes les observations sont répondantes. Notons $Y = \sum_r X_i$ le total de la variable d'intérêt dans l'échantillon.

L'estimateur non pondéré de la proportion p est $\tilde{p} = \frac{Y}{n}$. Sa variance asymptotique $\tilde{p}$ est estimée simplement par , $\hat{V} = \frac{\tilde{p}(1-\tilde{p})}{n}$ et le coefficient de variation associé par $\tilde{C} = \frac{\sqrt{\tilde{V}}}{\tilde{p}} = \sqrt{\frac{(1-\tilde{p})}{n\tilde{p}}}$.

Le total **X** correspondant est estimé de la même manière par $\tilde{X} = N\tilde{p} = \frac{N}{n} Y$, des estimateurs de la variance et du coefficient de variation sont donnés par :

$$V' = \left(\frac{N}{n}\right)^2 Y(1-\frac{Y}{n}) = \left(\frac{N}{n}\right)\tilde{X}(1-\frac{\tilde{X}}{N})$$ et

$$C' = \sqrt{\frac{1-Y/n}{Y}} = \sqrt{\frac{N(1-\frac{\tilde{X}}{N})}{n\tilde{X}}}\ .$$

On propose donc les estimateurs approchés suivants pour le coefficient de variation :

a) Total d'une variable indicatrice X

$$C_t = \sqrt{\frac{N(1-\frac{\hat{X}}{N})}{n\hat{X}}}$$ avec

$\hat{X}$	total estimé de la variable X dans la population
n	nombre de répondants à l'enquête
N	nombre de ménages estimé à l'enquête

On peut encore écrire $C_t = \sqrt{\frac{1-\hat{p}}{n\hat{p}}}$, où $\hat{p}$ est l'estimateur pondéré de la proportion de ménages possédant la caractéristique étudiée.

b) Proportion de ménages possédant une caractéristique donnée

La caractéristique étudiée est toujours décrite par l'indicatrice X.

$$C_p = \sqrt{\frac{(1-\hat{p})}{n\hat{p}}}$$ avec

n	nombre de ménages répondants à l'enquête
$\hat{p}$	estimateur pondéré de la proportion p dans la population

Cet estimateur est le même que celui du coefficient de variation du total de la variable X.

c) Proportion de ménages appartenant à un sous-champ de la population totale, possédant une caractéristique donnée

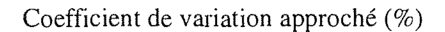
$$C_t' = \sqrt{\frac{(1-\hat{p})}{n_1\hat{p}}}$$

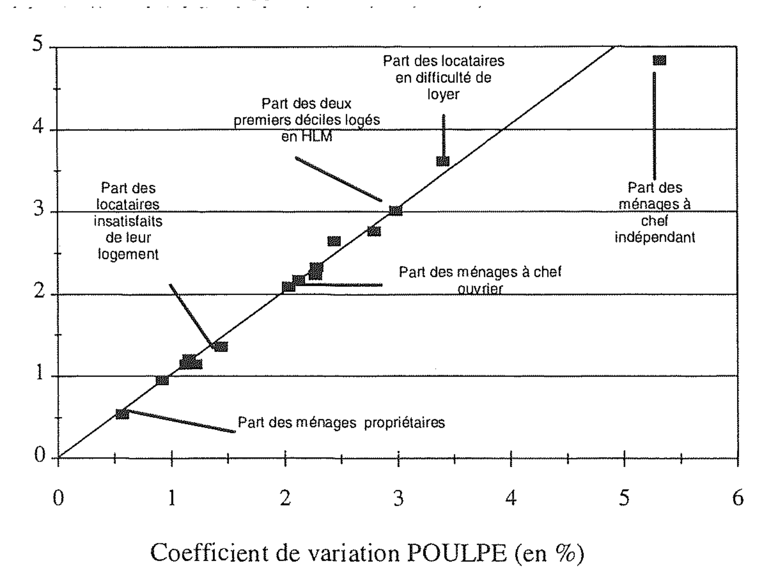
avec

n_1	nombre de ménages dans l'échantillon appartenant au sous-champ étudié
$\hat{p}$	estimateur pondéré de la proportion p dans la population

Les **graphiques 3 et 4** montrent que ces coefficients de variation approchés s'écartent très peu des valeurs calculées par POULPE pour les variables qui n'ont pas été calées. En revanche, pour les variables corrélées aux variables de calage, la précision donnée par POULPE est meilleure.

Graphique 3
Comparaison des coefficients de variation estimés par
avec des coefficients de variation approchés tirés des formules du II-2)
- Cas des proportions





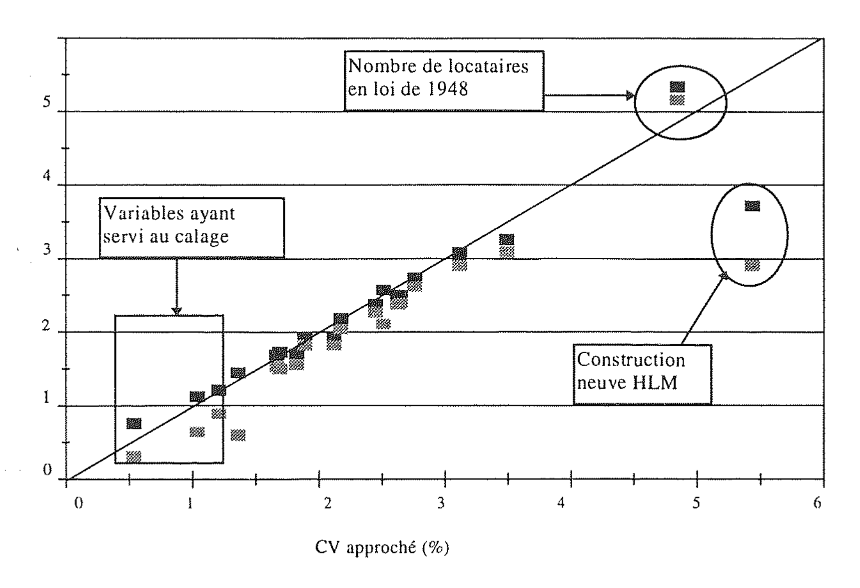
Ainsi, pour exploiter l'enquête Logement, POULPE n'est pas nécessaire pour estimer l'ordre de grandeur de la précision d'une variable indicatrice quelconque : une formule approchée donne un résultat satisfaisant. Ce résultat n'est sans doute vrai que parce que l'échantillon de l'enquête est suffisamment important ; il n'est certainement plus valable pour d'autres enquêtes dont l'échantillon est plus réduit.

2) Le calage sur marge améliore de façon importante la précision des variables corrélées aux variables ayant servi au calage. Sur les autres variables, le gain de précision est négligeable.

Les graphiques 1 et 2 montrent que les coefficients de variation calculés par POULPE avant et après le passage de CALMAR diffèrent très peu, lorsque la variable étudiée n'a pas servi au calage.

En revanche, pour les variables ayant servi au calage, le gain de précision dû à CALMAR est important. Le coefficient de variation après calage est nettement inférieur à la valeur avant calage et aussi à la valeur approchée.

Graphique 4
Comparaison des coefficients de variation estimés par POULPE, avant et après calage, avec des coefficients de variation approchés tirés des formules du II-2) - Cas des effectifs



De façon plus anecdotique, on note que pour certaines variables importantes, l'estimation de la proportion des ménages concernés est meilleure (au sens du coefficient de variation) que l'estimation des effectifs. Cela signifie que les erreurs commises au numérateur et au dénominateur sont positivement corrélées. L'exemple le plus net est donné par la proportion de ménages propriétaires, avec un coefficient de variation après calage de 0,17, contre 0,31 pour le nombre de propriétaires. Toutefois, pour la grande majorité des variables étudiées, les coefficients de variation du total et de la proportion sont très proches.

I-3) Efficacité du plan de sondage et limites des outils actuels

Les résultats produits par POULPE permettent de juger de l'efficacité du plan de sondage de l'enquête, en même temps que de la perte ou du gain de précision dus à l'échantillon-maître. Pour la plupart des variables, le design effect estimé par

POULPE est très proche de 1 : cela prouve que l'échantillon de l'enquête a été correctement tiré.

Quelques cas surprenants méritent d'être mentionnés :

- les propriétaires et la population des ménages. Avant calage, le « design effect » est très fort (plus de 2), même si le calage a pour effet de le ramener au voisinage de 1. Or, ces variables sont censées être bien couvertes par l'échantillon-maître ;

- tout ce qui concerne le secteur HLM (hors construction neuve). Le calage améliore de façon importante la précision des estimateurs ;

- les loyers moyens en HLM et dans le secteur privé. Le design effect calculé par POULPE est très faible (entre 0,2 et 0,3).

Par ailleurs, la modélisation actuelle du plan de sondage de la BSLN (par un sondage stratifié) se traduit par une sous-estimation de la variance pour toutes les variables fortement corrélées avec la construction neuve. L'illustration de cette limite est visible sur le **graphique 4**. La variable CNHLM, qui représente l'effectif des constructions HLM entre décembre 1993 et décembre 1996, est nettement séparée des autres variables.

Conclusion

La brève analyse présentée ici permet d'insister sur les apports du logiciel POULPE. Ces apports sont particulièrement nets sur deux plans :

- du point de vue méthodologique d'abord : POULPE permet de chiffrer le gain dû au calage sur les marges, technique maintenant largement employée pour redresser les enquêtes ; il permet aussi de juger d'un coup d'oeil la pertinence de l'échantillonnage d'une enquête, par l'examen des design effects ; il facilite enfin la détection de variables 'à problèmes', mal représentées par l'échantillon-maître ou la BSLN ;

- du point de vue du statisticien d'enquête ensuite. POULPE s'avère précieux pour donner des intervalles de confiances sur des fonctions de variables quantitatives, comme le revenu moyen ou les loyers moyens. Si, pour l'enquête Logement, des formules approchées permettent d'approcher correctement la précision d'une large classe d'estimateurs, cela vient sans doute de la taille importante de l'échantillon ; il n'est pas sûr que ce résultat soit valable pour d'autres enquêtes moins lourdes. POULPE permet aussi de juger de la pertinence d'une procédure de redressement, en quantifiant le gain de précision apporté par le calage sur des variables-clés d'une enquête.

La mise à disposition de POULPE, par les facilités qu'elle apporte, a le mérite de rappeler au statisticien d'enquête que la diffusion des résultats d'une enquête devrait systématiquement inclure des indications sur la précision de ces résultats. Les

estimations de précision permettent vis-à-vis des utilisateurs extérieurs de souligner qu'une enquête génère nécessairement des aléas et donc des intervalles de confiance sur les estimations ; elles se révèlent précieuses lorsque certaines évolutions entre deux enquêtes ne sont pas significatives. Les utilisateurs des enquêtes de l'Insee sont en effet plus enclins à s'interroger sur la précision des chiffres qu'il y a quelques années.

Bibliographie indicative

CARON, N. : « Calcul de l'effet de sondage dans le logiciel POULPE », note interne n° 981/F410, 1996.

DEVILLE, J.C. : « Estimation de la précision de données d'enquêtes », document de travail de la Direction des Statistiques Démographiques et Sociales, n° F 9211, Insee, 1992.

DEVILLE, J.C., CARON, N., SAUTORY, O., « Estimation de la précision de données d'enquêtes : document méthodologique sur le logiciel POULPE », document de travail de l'Unité de Méthodologie Statistique, à paraître.

LACROIX, T. : « Pondérations de l'enquête Logement 1992/1993 et révision des pondérations des enquêtes logement 1984 et 1988 », document de travail de la Direction des Statistiques Démographiques et Sociales, n° F 9408, Insee, 1994.

LAFERRERE, A. : « Les ménages et leurs logements », *Insee Première* n° 562, Insee, 1997.

SÄRNDAL, C.E., SWENSSON, B., WRETMAN, J.: *Model assisted Survey Sampling*, Springer-Verlag, 1992

Annexe 1

Variables traitées
Précision et intervalle de confiance à 95 %

Totaux

Variable	Valeur (milliers)	Ecart-type CALMAR	borne inférieure (milliers)	borne supérieure (milliers)
Population des ménages	57785	134773	57521	58049
Nombre de ménages dont la personne de référence a plus de 65 ans	5642	36113	5571	5713
Nombre de propriétaires	12645	39016	12569	12721
Nombre de locataires HLM	3657	21805	3614	3700
Nombre de locataires privés	4449	39447	4372	4526
Nombre de bailleurs privés	1579	32101	1516	1642
Nombre de logements mis en location par des bailleurs privés	2959	124661	2715	3203
Nombre de ménages possédant une résidence secondaire	2051	37555	1977	2125
Nombre de locataires insatisfaits de leur logement	1007	26373	955	1059
Nombre de ménages en loi de 1948	337	17403	303	371
Nombre de ménages dont la personne de référence a entre 20 et 23 ans	640	19759	601	679
Nombre de ménages dont la personne de référence a entre 24 et 27 ans	1264	28703	1208	1320
Nombre de locataires sans bail	797	23147	752	842
Nombre de locataires des deux premiers déciles	2507	37661	2433	2581
Nombre de locataires appartenant aux deux premiers déciles de revenu logés en HLM	1204	25323	1154	1254
Nombre de locataires appartenant aux deux premiers déciles de revenu logés par un bailleur privé	1110	26400	1058	1162
Nombre de ménages à chef indépendant	1092	26143	1041	1143
Nombre de nouveaux ménages	2185	34276	2118	2252
Construction neuve HLM	268	7789	253	283
Nombre d'accédants récents	1658	30222	1599	1717
Nombre de ménages ayant changé de logement dans la même commune depuis 1992	2570	39599	2492	2648

Ratios

Variable	Valeur	Ecart-type CALMAR	borne inférieure	borne supérieure
Revenu mensuel moyen déclaré	13 050	51	12 951	13 149
Prix moyen d'acquisition des logements	669 440	9363	651 089	687 791
Loyer moyen en HLM	1 679	7.6	1 664	1 694
Loyer moyen dans le parc locatif privé	2 517	14.6	2 488	2 546
Nombre moyen de logements mis en location par les bailleurs privés	1.87	7.05E-02	1.73	2.01
Nombre de personnes par ménage	2.48	5.77E-03	2.47	2.49
Part des ménages propriétaires	54.3	0.167	54.0	54.6
Part des ménages insatisfaits	6.0	0.133	5.7	6.3
Part des ménages locataires HLM	15.7	0.093	15.5	15.9
Part des ménages locataires privés	19.1	0.169	18.8	19.4
Part des ménages à chef ouvrier	20.9	0.212	20.5	21.3
Part des ménages bailleurs privés	6.8	0.137	6.5	7.1
Part des locataires logés en HLM	41.1	0.264	40.6	41.6
Part des locataires en locatif privé	50.1	0.340	49.4	50.8
Part des locataires des deux premiers décites de revenu logés en HLM	25.2	0.480	24.3	26.1
Part des locataires sans bail	9.0	0.258	8.5	9.5
Part des locataires des deux premiers déciles de revenu en difficulté de loyer	29.5	0.820	27.9	31.1
Part des locataires en difficulté de loyer	17.1	0.344	16.4	17.8
Part des ménages sans confort trouvant leurs conditions de logement insuffisantes	12.0	0.400	11.2	12.8
Part des ménages à chef indépendant	4.7	0.112	4.5	4.9
Part des ménages locataires en loi de 1948	1.5	0.074	1.3	1.6

Annexe 2

Application numérique des formules approchées

On cherche à évaluer la précision de trois variables : nombre de locataires du parc privé, part des ménages logés dans le parc locatif privé, part des locataires logés dans le parc privé. Les données tirées de l'enquête sont les suivantes :

Nombre de ménages répondants	29043
Nombre de ménages locataires dans l'échantillon	11096
Nombre de ménages estimé	23 286 000
Nombre de locataires estimé	8 877 000
Nombre de locataires du parc privé estimé	4 449 000
Estimation de la part des ménages logés dans le parc privé	19.1 %
Estimation de la part des locataires logés dans le parc privé	50.1 %

D'où les coefficients de variation approchés suivants :

Quantité	Coefficient de variation approché
nombre de locataires du parc privé	$C_1 = \sqrt{\frac{23286000}{29043}\left(\frac{1-4449000/23286000}{4449000}\right)} = 1.21$
part des ménages logés dans le parc locatif privé	$C_2 = \sqrt{\frac{(1-0.191)}{(29043)(0.191)}} = 1.21$
part des locataires logés dans le parc privé	$C_3 = \sqrt{\frac{(1-0.501)}{(11096)(0.501)}} = 0.95$

On vérifie bien que les coefficients de variation approchés pour le nombre de locataires privés et la part des ménages locataires privés sont les mêmes.

Session 4

Les indices

ÉTUDE DU CHAÎNAGE D'INDICES DE PRIX À L'AIDE DE MICRO-DONNÉES

F. Magnien et J. Pougnard

1. Introduction

Les substitutions que les consommateurs effectuent entre produits ou entre lieux d'achats constituent l'une des difficultés majeures de la construction d'indices de prix. Comment, en effet, mesurer entre deux dates des évolutions différentes - voire divergentes - de prix, tout en prenant en compte les substitutions induites ? Il faut, d'une façon ou d'une autre, agréger les prix au moyen des quantités vendues alors que ces quantités ne cessent d'évoluer. Le plus simple consiste à les fixer. On obtient ainsi les indices de Laspeyres et de Paasche directs suivant que sont retenues les quantités initiales ou bien finales :

$$L_{T/0}^{D} = \frac{\sum_{s} q_0^s p_T^s}{\sum_{s} q_0^s p_0^s} = \sum_{s} w_0^s \frac{p_T^s}{p_0^s} \text{ où } w_0^s = \frac{q_0^s p_0^s}{\sum_{s'} q_0^{s'} p_0^{s'}}$$

$$P_{T/0}^{D} = \frac{\sum_{s} q_T^s p_T^s}{\sum_{s} q_T^s p_0^s} = \left[\sum_{s} w_T^s \left(\frac{p_T^s}{p_0^s} \right)^{-1} \right]^{-1} \text{ où } w_T^s = \frac{q_T^s p_T^s}{\sum_{s'} q_T^{s'} p_T^{s'}}$$

Dans ces expressions, *s* désigne un produit dans un point de vente (une « série »), 0 la période initiale, dite de « base » et *T* la période finale (sous revue). Malgré leur mise en oeuvre particulièrement simple, ces indices ont rapidement montré leurs limites sur longue période. En effet, ils interdisent catégoriquement la prise en compte des substitutions.

Une autre approche consiste à agréger les prix à l'aide des quantités courantes. Elle conduit à la notion d'indice de prix moyen, encore appelé indice de "valeurs unitaires", apprécié des professionnels :

$$M_{T/0} = \frac{\sum_s q_T^s p_T^s / \sum_s q_T^s}{\sum_s q_0^s p_0^s / \sum_s q_0^s}$$

La critique immédiate soulevée par ce type d'indice est la suivante : si des substitutions s'opèrent des points de vente les plus chers vers les moins chers bien que les prix restent fixes, l'indice baissera. Ainsi, alors que l'on reproche aux indices de Laspeyres et de Paasche de ne pas prendre en compte les substitutions, c'est la critique inverse qui est faite aux indices de valeurs unitaires. A. Saglio (1995) et W. J. Hawkes (1995) ont analysé l'écart entre indice de Laspeyres et indice de prix moyen.

La théorie économique n'a certes pas encore trouvé l'indice idéal ; elle propose cependant une analyse satisfaisante de la prise en compte des substitutions : l'approche dite de l'*indice à utilité constante* (IUC). L'IUC[1] est le rapport de deux dépenses : la dépense initiale du consommateur et une dépense courante fictive celle qui, compte tenu des changements de prix, assure au moindre coût, par des substitutions appropriées entre produits ou points de vente, un niveau d'utilité à la période courante égal à ce qu'il était initialement. L'écart entre l'indice de Laspeyres direct et l'IUC est souvent (ce sera le cas ici) appelé "biais de substitution".

La mise en oeuvre de l'IUC est délicate puisqu'elle suppose connue la fonction d'utilité des consommateurs. Deux méthodes ont été développées. La première, que l'on peut qualifier de *paramétrique*, consiste à faire l'hypothèse que cette fonction d'utilité est d'une forme spécifiée : quadratique, Cobb-Douglas, ... et à considérer l'IUC associé, dit "exact" pour la forme d'utilité retenue (suivant la terminologie de Diewert (1976)). On montre ainsi que l'indice de Fisher :

$$F_{T/0}^D = \sqrt{L_{T/0}^D P_{T/0}^D}$$

est exact pour la fonction d'utilité quadratique homogène ; l'indice de Törnqvist

$$T_{T/0}^D = \prod_s \left(\frac{p_T^s}{p_0^s} \right)^{\frac{w_0^s + w_T^s}{2}}$$

est exact pour la fonction de coût translog, alors que la moyenne géométrique pondérée :

1. Cost-of-living (COL) index en anglais.

$$G_{T/0}^{D} = \prod_{s} \left(\frac{p_T^s}{p_0^s} \right)^{w_0^s}$$

est exacte pour la fonction d'utilité de Cobb-Douglas (Cf. Diewert (1976)). De fait, la formule géométrique simple :

$$GS_{T/0} = \prod_{s} \left(\frac{p_T^s}{p_0^s} \right)^{1/n}$$

où n désigne le nombre de séries, est utilisée en France dans le calcul de l'indice des prix à la consommation (IPC) pour certaines variétés[2]. Les indices de Fisher et de Törnqvist sont considérés comme de bonnes approximations de l'IUC : c'est souvent par rapport à eux que l'on mesure le biais de substitution de l'indice de Laspeyres.

La seconde méthode utilisée pour le calcul de l'IUC, dite non paramétrique, s'appuie sur la "théorie des préférences révélées" développée par Afriat (1967)[3] : plutôt que de se donner une fonction d'utilité d'une forme particulière à partir de laquelle on obtient une expression analytique de l'IUC, on infère cette fonction d'utilité des données (prix et quantités) en exploitant l'hypothèse sous-jacente à l'IUC selon laquelle les agents sont rationnels, c'est à dire cherchent à atteindre un niveau d'utilité donné au moindre coût.

Une approche du problème des substitutions alternative à celle de l'IUC est possible *: le chaînage d'indices.* Le principe consiste à découper le temps en intervalles de longueurs égales et à remettre à jour les pondérations au début de chacune de ces périodes ; à l'intérieur des périodes, on procède au calcul d'un indice direct :

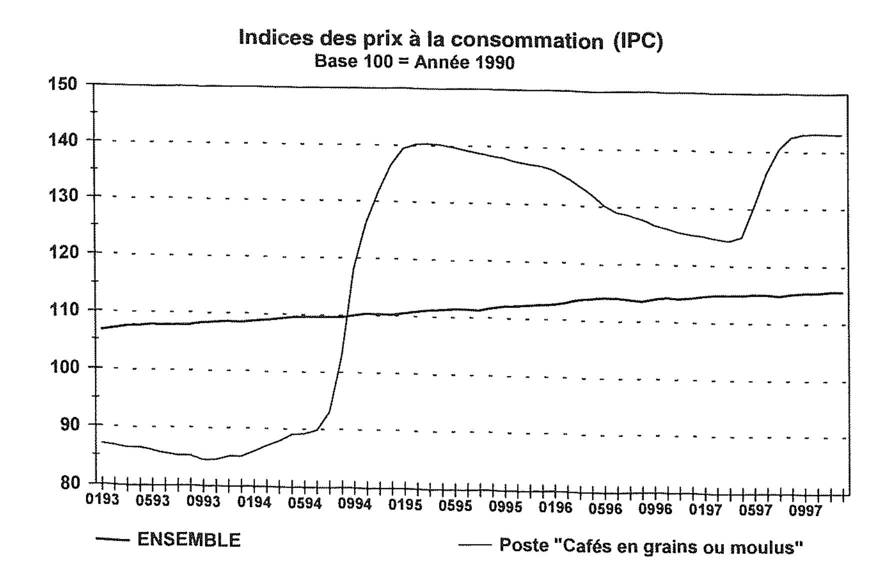
$$I_{T/0}^{C} = \prod_{t=1}^{T} I_{t/t-1}^{D}$$

L'une des questions fondamentales dans le chaînage d'indices est le choix de la durée séparant deux chaînages successifs. Alors que les États-Unis attendent plus de dix ans pour réviser leurs pondérations et que les Allemands chaînent leur indice tous les cinq ans, la France le fait tous les ans. On peut penser que cette durée doit être la plus courte possible. Le recours progressif à des micro-données (scanner data), dont l'un des avantages est de fournir avec une fréquence élevée (elle peut être hebdomadaire) les quantités en plus des prix, permet d'envisager de réduire encore la durée séparant deux chaînages successifs, et d'approcher ainsi l'indice de Divisia[4].

2. Pour le calcul mensuel de l'IPC, on ne dispose pas des quantités mais seulement des prix.
3. Voir aussi Diewert (1973).
4. Rappelons - résultat classique - qu'avec une fonction d'utilité homothétique, l'IUC coïncide avec l'indice de Divisia (Cf. Hulten (1973), Reinsdorf (1998)).

Certains travaux, notamment de B. Szulc (1983), laissent penser qu'en période de prix chahutés, un chaînage trop fréquent peut aller à l'encontre d'une meilleure prise en compte des substitutions. Szulc exprime mathématiquement l'écart à l'unité du ratio "Laspeyres chaîné/Laspeyres direct" comme une accumulation de taux (propres à chaque période de chaînage) dont les signes dépendent d'une part des substitutions opérées par les consommateurs et, d'autre part, du "rebond" des prix.

Des micro-données fournies par la société AC Nielsen nous ont permis de tester cette analyse. Ces données portent sur le café[5] pour une période, 1994-1996, durant laquelle les cours ont connu de très fortes variations :



En effet, l'année 1994 a été le théâtre d'une flambée des cours par suite de gelées au Brésil[6], premier producteur mondial. En 1995, les cours se sont stabilisés puis ont baissé régulièrement en raison notamment d'une réduction générale des stocks. En 1996, la conjonction d'une mauvaise récolte et d'une restriction des exportations chez les deux principaux producteurs, le Brésil et la Colombie, a enrayé cette baisse.

L'utilisation de données scannées est complexe. Ces données contiennent en effet une masse très importante d'informations : outre les prix et les quantités, une description extrêmement détaillée des points de vente et surtout des produits est fournie. Il en résulte un nombre considérable de séries : environ 30 000 pour le café contre 500 seulement dans l'IPC français !

5. D'autres études ont été menées à partir de données scannées sur le café : Haan-Opperdoes (1997a), Hawkes (1995), Reinsdorf (1995) notamment.
6. Phénomène amplifié par le mécanisme du marché à terme sur lequel s'échange la production mondiale.

Le problème des séries manquantes devient alors aigu. A priori, la multiplicité des séries manquantes peut être imputée à la multiplicité des produits suivis : relativement, le nombre des points de vente de la base AC Nielsen n'est guère plus élevé que dans l'IPC. Il s'avère que c'est la dispersion entre points de vente des taux de distribution des produits[7] qui explique d'abord la disparition de séries. En outre, la dispersion des prix résulte largement de la diversité des produits, de sorte que c'est d'abord entre produits que les substitutions ont lieu (il convient de préciser ici que le panel AC Nielsen utilisé dans cette étude ne contient que des hypermarchés et des supermarchés). Ces considérations nous ont amenés à résoudre le problème des séries manquantes par agrégation des lieux d'achats selon quatre formes de vente. Les données, hebdomadaires, ont également été agrégées mensuellement.

Le calcul des indices définis précédemment a pu ainsi être mené pour le "poste" café et pour chacune des "variétés" qui le constituent. Les résultats obtenus mettent en lumière le phénomène de dérive du chaînage :

(*i*) pour le poste et la plupart des variétés, une dérive régulière des indices chaînés par rapport aux indices directs, l'indice de Laspeyres (de Paasche) chaîné majorant (minorant) l'indice direct ;

(*ii*) cette dérive, légère avec un chaînage annuel, devient importante avec un chaînage mensuel ; à plusieurs reprises, l'indice chaîné mensuellement *augmente* quand l'indice direct *diminue* (phénomène inverse avec l'indice de Paasche).

Il était naturel de soumettre ces faits, confirmés par d'autres études récentes[8], au crible de l'analyse proposée par B. Szulc. Au niveau du poste, les résultats sont partagés : l'explication de la dérive des indices chaînés par une corrélation négative entre la variation courante des prix et leur variation depuis la période de base (le phénomène de « rebond ») n'est vraiment acceptable qu'en 1996 ; toutefois, pour certaines variétés, l'adéquation des résultats obtenus avec l'analyse de Szulc s'avère excellente, dans le cas du café moulu décaféiné en 1996 notamment.

L'organisation de la suite de cette étude est la suivante :

2. Description du panel AC Nielsen sur le café
3. L'instabilité des séries
4. Gestion des séries manquantes
5. Calculs d'indices directs et chaînés
6. Analyse de la dérive du chaînage
7. Conclusion

---

7. On entend par là la proportion de l'ensemble des produits qui sont disponibles dans un point de vente.
8. Cf. note de bas de page 5.

2. Description du panel AC Nielsen[9]

La base de données sur laquelle s'appuie cette étude est une extraction du panel SCANTRACK de la société AC Nielsen relative au café torréfié. Elle contient, pour un *échantillon* important d'hypermarchés et supermarchés, et pour la quasi-totalité des produits offerts sur le marché, les *chiffres d'affaires* et les *quantités* vendues chaque semaine d'une période allant de janvier 1994 à décembre 1996.

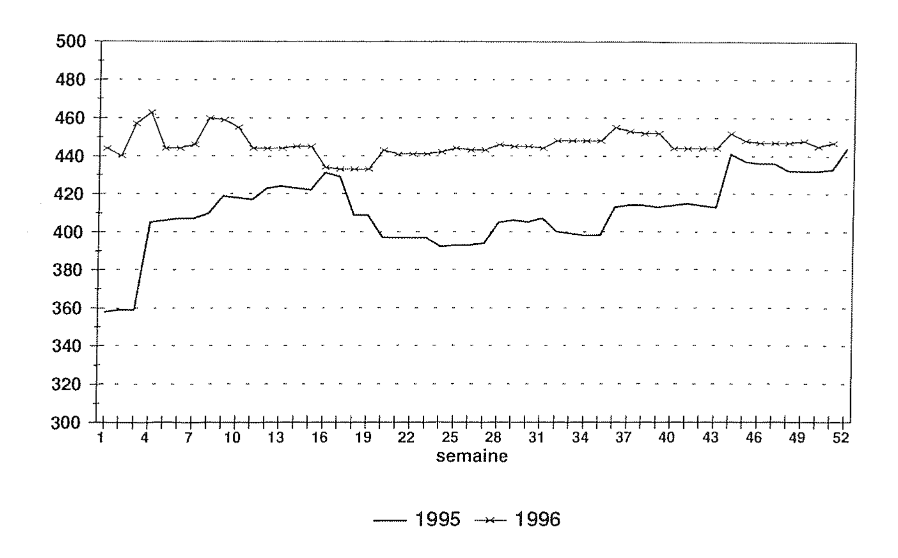
2.1. Les points de ventes

Dans le panel AC Nielsen, un point de vente est décrit à l'aide de trois critères : la *forme de vente*, l'*enseigne* et la *localisation*. Les seules formes de vente retenues sont les hypermarchés et les supermarchés. Les hypermarchés de moins de 6 500m^2 (HM-) ont été différenciés des autres (HM+) ; de même, les supermarchés de moins de 1 200m^2 (SM-) ont été distingués des autres (SM+). Bien que retenues par AC. Nielsen comme critère de stratification pour le tirage de leur échantillon de points de vente, les enseignes auxquelles appartiennent les points de vente ne nous ont pas été communiquées. Quant à leur localisation géographique, seuls les départements dans lesquels sont situés les points de vente nous ont été notifiés. La localisation des points de vente n'intervient pas dans cette étude.

Sur la période 1994-1996, 575 points de ventes différents ont fait partie au moins une fois du panel. En tendance, le nombre de points de vente n'a cessé d'augmenter, surtout en 1994, première année de suivi du panel. En 1996, l'effectif des points de vente s'est stablilisé, autour de 440 (graphique 1).

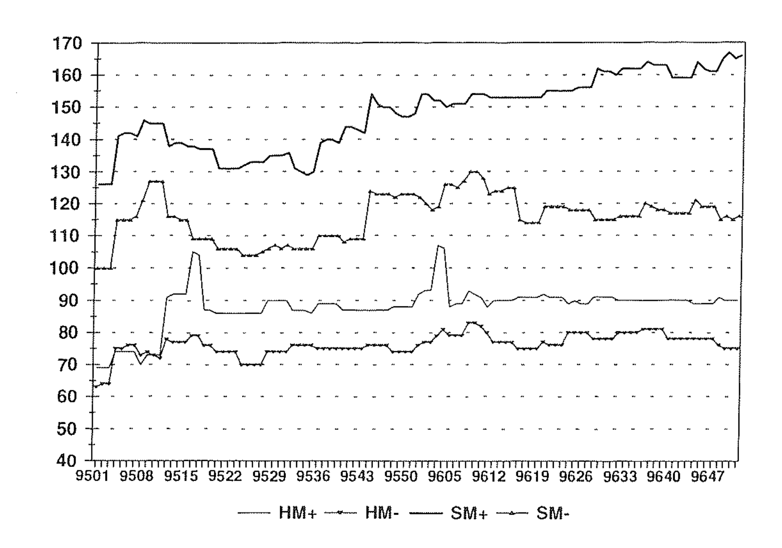
9. Utilisé dans cette étude.

Graphique 1 : Evolution du nombre de points de vente



Par forme de vente, on retrouve les mêmes évolutions. Cependant, le nombre de supermarchés a augmenté plus fortement, notamment les plus grands (SM+).

Graphique 2 : Evolution du nombre de points de vente par forme de vente



La répartition par formes de vente du chiffre d'affaires ("extrapolé" - cf. Section 2.3.) du café est restée stable au cours des trois années : environ 27 à 28 % ont été réalisés par les grands hypermarchés, 20 % par les plus petits, 35 % par les grands supermarchés et entre 15 et 16 % par les autres.

2.2. Les produits élémentaires

Un *produit élémentaire* est défini de façon extrêmement fine par le croisement d'un ensemble de caractéristiques. Les caractéristiques retenues dans cette étude[10] sont au nombre de 10, regroupées, pour la présentation ci-dessous en 3 catégories :

Production et commercialisation :

1. Fabricant (ou distributeur)
2. Marque
3. Référence

Conditionnement :

4. Type d'emballage d'un paquet (bocal, boîte, carton, ...)
5. Nombre de paquets vendus ensemble
6. Poids de l'ensemble

Description du "contenu" :

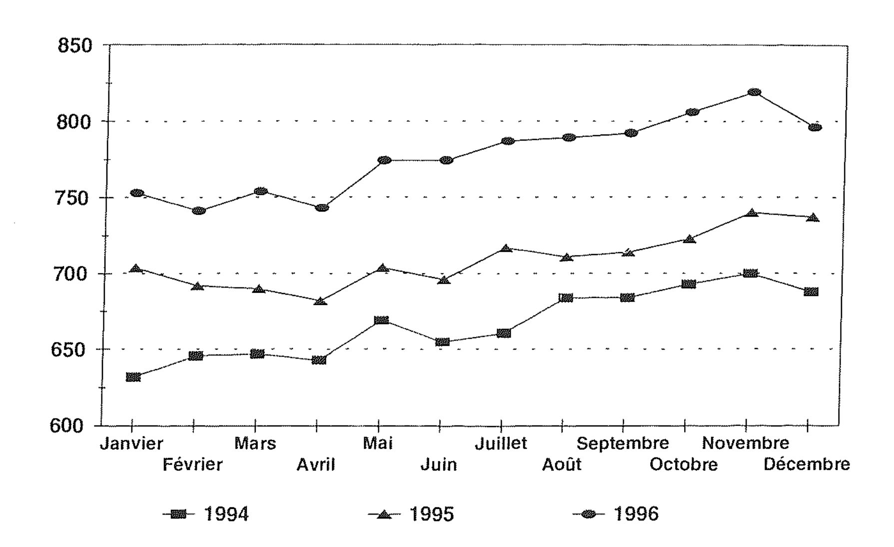
7. Le type (grain, moulu normal, moulu expresso)
8. La qualité (normal, décaféiné)
9. La gamme (arabica, robusta, mélange)
10. L'origine (Brésil, Colombie, ...)

A chacune de ces 10 caratéristiques est associé un certain nombre de modalités (de 2 pour la qualité à 30 pour l'origine). Il s'ensuit que le nombre de produits élémentaires envisageables est très élevé. Seule une partie de ces produits est fabriquée et commercialisée : près de 1 200 produits élémentaires différents ont appartenu au panel au moins une fois au cours des trois années. Ces produits élémentaires se répartissent entre les différents types de café : 871 pour le café moulu normal, 121 pour le café moulus expresso et 176 pour le café en grains. Selon la gamme, il y a 709 arabica, 381 mélanges et 78 robusta.

---

10. Les autres caractéristiques sont relatives au suivi technique du panel. Elles ne sont pas liées à la nature du produit. Elles n'ont donc pas été prises en considération, grâce à un travail d'agrégation préalable (cf. Section 4).

Graphique 3 : Evolution du nombre de produits élémentaires
(nombre moyen présent par mois)



Le nombre de produits élémentaires suivis dans le panel a cru régulièrement, passant de 630 en janvier 1994 à 800 en décembre 1996.

Dans l'IPC, on regroupe les produits élémentaires par *variétés*. On calcule alors des indices de prix pour chaque variété puis on agrège ces indices pour obtenir les indices des *postes*. Il est en effet intéressant, pour l'analyse de l'évolution des prix, de disposer d'indice pour des regroupements étroits de produits à l'intérieur de chaque poste ; en outre, l'absence de relevés de quantités dans l'IPC *nécessite* de passer par ce niveau et même par le niveau plus élémentaire encore des « variété-agglomération ». Avec les micro-données, cette contrainte technique ne se pose pas : le calcul d'indice par variété n'a alors pour objet que de mettre en évidence des évolutions de prix différentes à *l'intérieur* d'un poste.

Le poste « café en grains ou moulu » de l'IPC est représenté par trois variétés. La richesse du panel AC Nielsen permet d'en introduire deux autres. Au total, les variétés suivantes ont été retenues :

(1) Café moulu normal, non décaféiné, robusta ou mélange
(2) Café moulu normal, décaféiné
(3) Café moulu normal, non décaféiné, arabica
(4) Café en grains
(5) Café moulu expresso

Par exemple, la variété (1) est constituée des produits élémentaires qui prennent la modalité "moulu normal" de la caractéristique 7, la modalité "non décaféiné" de la

caractéristique 8 et l'une ou l'autre des modalités "robusta" ou "mélange" de la caractéristique 9, les modalités des autres caractéristiques étant quelconques.

Tableau 1 : Pondération annuelle[(1)] des variétés selon leur chiffre d'affaires

	1994	1995	1996
Variété 1	24,6	28,1	25,8
Variété 2	7,1	7,1	7,3
Variété 3	50,5	48,4	50,0
Variété 4	4,8	4,3	4,1
Variété 5	13,0	12,1	12,8
Poste	100	100	100

(1) En %

Dans l'IPC, les variétés[11] sont classées selon deux catégories : les *homogènes* et les *hétérogènes*. Les variétés homogènes réunissent des produits très proches par leurs caractéristiques, donc de prix voisins. Il en est ainsi des variétés 1 et 3. Les autres (2, 4 et 5) sont hétérogènes. L'absence de relevés de quantités conduit à utiliser des formules de calcul différentes (au niveau des agglomérations), selon que l'on a affaire à une variété homogène ou hétérogène. Cette distinction n'est pas nécessaire pour le calcul d'indices dès lors que l'on dispose, comme c'est le cas avec le panel AC Nielsen, de relevés de quantités.

2.3. *Quantités et prix*

Chaque semaine t, pour chaque couple $s = (i, j)$, où i désigne un produit élémentaire et j un point de vente, sont relevés une quantité q_t^s et un chiffre d'affaires ca_t^s. Le panel contient, en outre, des *coefficients d'extrapolation* λ_t^s permettant, pour chaque produit élémentaire i et chaque semaine t, d'estimer par

$$\sum_j \lambda_t^{i,j} q_t^{i,j} \quad \text{et} \quad \sum_j \lambda_t^{i,j} ca_t^{i,j}$$

la quantité vendue et le chiffre d'affaires réalisé dans l'*univers* des points de vente (dans les expressions précédentes, j parcourt l'*échantillon* des points de vente de la semaine t). Intuitivement, l'application de ces coefficients d'extrapolation revient à

11. dites "ordinaires" ; il y a aussi les *produits frais*, les *biens durables* et les *tarifs*. Le poste « café » de l'IPC ne contient que des variétés ordinaires.

considérer qu'à la semaine t, chaque point de vente j du panel représente un ensemble de λ_t^s ($s = (i, j)$) points de vente pour la vente du produit i.

En fait, ce coefficient d'extrapolation est le même pour tous les produits élémentaires, ne dépend des points de vente que par les "strates" auxquelles ils appartiennent et des semaines que par les mois dont elles font partie. En effet, le panel AC Nielsen résulte d'un sondage aléatoire stratifié des points de ventes (avec allocation optimale au sens de Neyman à l'intérieur de chaque strate). Les critères de stratification retenus sont la taille de la forme de vente (HM+, HM-, SM+, SM-), l'enseigne, la localisation régionale et la taille de l'agglomération.

Pour un produit élémentaire i, la quantité vendue au cours d'une semaine t d'une année donnée A dans l'ensemble (et non l'échantillon) des points de vente j d'une strate J est

$$Q_t^i(J) = \sum_{j \in J} q_t^{i,j}$$

Comme le chiffre d'affaires annuel pour l'ensemble (exhaustif) des points de vente j de la strate J et de l'ensemble des produits i' (pas seulement les produits élémentaires du café) :

$$\sum_{t' \in A} \sum_{i'} \sum_{j \in J} q_{t'}^{i',j} p_{t'}^{i',j}$$

est connu, la quantité $Q_t^i(J)$ peut être estimée suivant la méthode du ratio par :

$$\sum_{j \in S_t(J)} \frac{\sum_{t' \in A} \sum_{i'} \sum_{j \in J} q_{t'}^{i',j} p_{t'}^{i',j}}{\sum_{t' \in A} \sum_{i'} \sum_{j \in S_t(J)} q_{t'}^{i',j} p_{t'}^{i',j}} q_t^{i,j}$$

où $S_t(J)$ est l'échantillon des points de vente de la strate J suivi la semaine t. Ainsi, le coefficient d'extrapolation pour une année donnée

$$\lambda_t^J = \frac{\sum_{t' \in A} \sum_{i'} \sum_{j \in J} q_{t'}^{i',j} p_{t'}^{i',j}}{\sum_{t' \in A} \sum_{i'} \sum_{j \in S_t(J)} q_{t'}^{i',j} p_{t'}^{i',j}}$$

ne dépend pas des produits et ne dépend que des strates auxquelles les points de vente appartiennent. En outre, l'échantillon $S_t(J)$ n'est révisé que toutes les quatre

semaines : il ne dépend que du mois m auquel t appartient (on écrira donc $S_m(J)$ et λ^J_m).

Par forme de vente, le nombre de points de vente que contient le panel diffère de ce que l'on observe dans l'univers des points de vente : les formes de vente les plus petites sont sous-représentées (il y avait réellement 347 HM+, 683 HM-, 2 784 SM+ et 3 887 SM- en janvier 1996 contre respectivement 97, 80, 152 et 122 dans le panel). Les multiplicateurs corrigent cette distorsion puisqu'ils sont d'autant plus grands que la taille de la forme de vente est petite.

Dans cette étude, il a été décidé de travailler sur des données mensuelles. On retrouve ainsi le rythme de calcul de l'IPC et de révision de l'échantillon des points de vente du panel AC Nielsen. Ceci permet en outre d'avoir une base de données moins lourde et donc plus maniable. Surtout, le problème de la perte de « séries » (cf. Section 3) s'en trouve notablement réduit. Ainsi, pour chaque couple $s = (i, j)$ constitué d'un produit élémentaire i et d'un point de vente j, le nombre d'unités vendues et le chiffre d'affaires des quatre semaines d'un même mois m ont été agrégés, ce qui a permis de calculer un prix mensuel moyen :

$$p^s_m = \frac{\sum_{t \in m} ca^s_t}{\sum_{t \in m} q^s_t} = \frac{ca^s_m}{q^s_m} \qquad (2.3.)$$

Ce prix est *unitaire* (les statisticiens parlent de "valeurs unitaires"). Or, suivant les modalités de la caractéristique 6 d'un produit élémentaire, la quantité (exprimée en grammes) de café contenue dans une *unité vendue* en magasin est variable d'un produit à un autre. D'où le choix d'une *unité de compte* commune : 250 grammes. C'est relativement à cette unité de compte que les prix et quantités seront dorénavant mesurés.

Si l'on revient un instant sur les formules de calcul des différents indices présentées dans l'introduction, il apparaît que toutes sont fondées sur des ratios p^S_T / p^S_0 où p^S_m ($m = 0, T$) est calculé selon 2.3. Il s'agit donc d'indices de valeurs unitaires. Balk (1998) a montré que ces indices sont très proches de l'IUC si les préférences sont représentées par une fonction d'utilité simplement additive (i.e. égale à la somme des quantités consommées). Cette hypothèse est raisonnable au niveau de désagrégation des produits élémentaires (si les quantités consommées d'une semaine à l'autre n'évoluent pas trop).

2.4. Les promotions

Le café est évidemment l'objet de *promotions*[12] ; c'est, entre autre, par ces promotions que la baisse des prix, amorcée en début d'année 1995, a été répercutée aux prix de détail. Les promotions prennent essentiellement la forme de quantités supplémentaires ou de réductions de prix.

Lorsqu'un produit élémentaire i est l'objet, une semaine t dans un point de vente j, d'une promotion, les modalités des caractéristiques 5 et 6 restent *inchangées* : ce sont les modalités hors promotion. Les promotions ne se traduisent donc jamais par l'apparition d'un nouveau produit élémentaire. Lorsqu'une promotion prend la forme d'une quantité supplémentaire offerte, celle-ci entraîne[13] une augmentation de la quantité q_t^s. Une promotion offrant une réduction de prix induit[14] une baisse du chiffre d'affaires ca_t^s.

3. L'instabilité des séries

3.1. La notion de série

Les Sections 2.3 et 2.4 ont mis en évidence l'importance des couples formés d'un produit élémentaire et d'un point de vente. Un tel couple s'appelle une *série*. Il s'agit d'un concept essentiel dans la construction d'indices de prix. Dans le panel AC Nielsen relatif au café, il y avait 26 770 séries en décembre 1994, 34 930 en décembre 1995 et 35 683 en décembre 1996. Ce nombre est considérable puisqu'il n'était que de l'ordre de 500 dans le poste « café » de l'IPC aux mêmes dates. La notion de série recouvre les deux types de substitutions auxquelles procèdent les consommateurs en réaction aux variations de prix, à l'absence des produits ou à d'autres facteurs (campagnes publicitaires, ...) : changements de produit et changements de lieu d'achat. La répartition des séries entre les cinq variétés du poste café est donnée dans le tableau suivant :

12. Une promotion est nécessairement indiquée en rayon.
13. En supposant le nombre d'achats inchangé.
14. Cf. note précédente.

Tableau 2 : Répartition des séries entre variétés

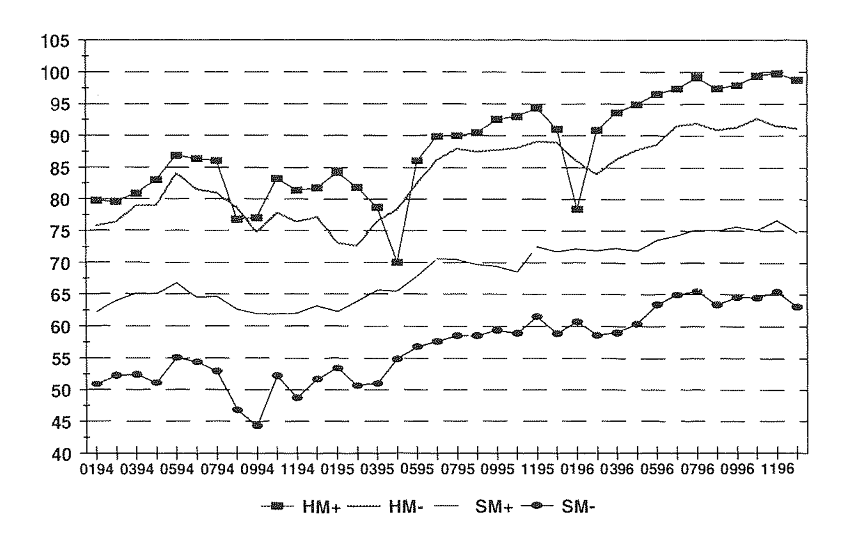
	Décembre 1994		Décembre 1995		Décembre 1996	
	Nombre	**%**	**Nombre**	**%**	**Nombre**	**%**
Variété 1	4 845	18,1	5 799	16,6	6 221	17,4
Variété 2	3 332	12,4	4 367	12,5	4 271	12,0
Variété 3	11 462	42,8	15 598	44,7	16 068	45,0
Variété 4	2 647	9,9	3 029	8,7	2 909	8,2
Variété 5	4 484	16,8	6 137	17,6	6 214	17,4
Poste	26 770	100	34 930	100	35 683	100

La variété 3 regroupe à elle seule près de la moitié du nombre des séries. Ceci reflète sa part dans le chiffre d'affaires du poste (cf. tableau 1).

3.2. Evolution du nombre de séries

La forte augmentation du nombre de séries résulte à la fois de l'accroissement du nombre de points de vente et du nombre de produits élémentaires suivis dans le panel (graphiques 1 et 3). L'augmentation du nombre moyen de produits élémentaires par point de vente s'observe quel que soit la forme de vente :

Graphique 4 : Evolution du nombre moyen de séries par point de vente, selon la forme de vente

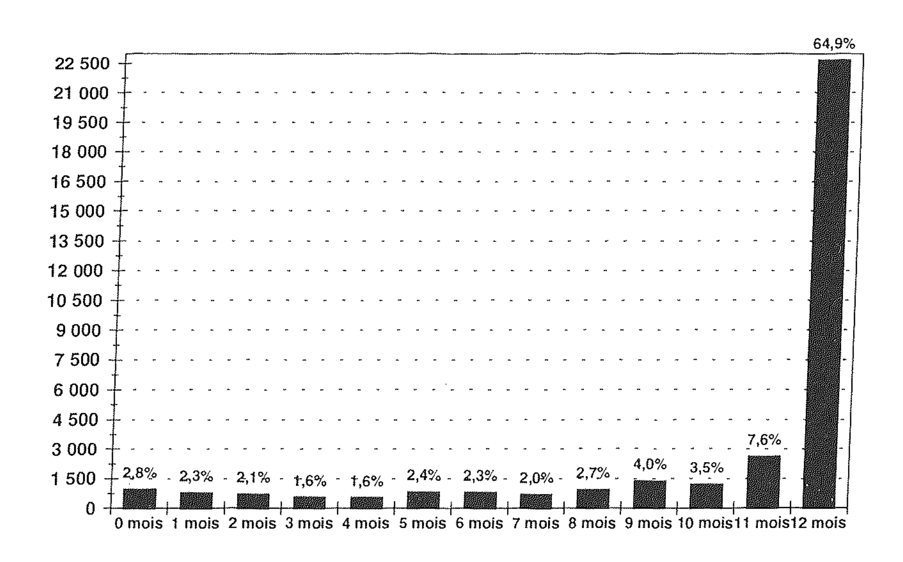


Le graphique 4 est relatif au poste « café » dans son ensemble. A l'intérieur de chaque variété, le nombre moyen de produits élémentaires par point de vente est d'autant plus élevé que la taille de la forme de vente est importante (toutefois, les formes HM+ et HM- ont à peu près le même nombre de séries). Le nombre moyen de séries par produit élémentaire augmente moins vite ; c'est donc d'abord l'élargissement de la gamme de produits offerts qui explique l'accroissement du nombre de séries.

3.3. *Instabilité des séries dans l'échantillon*

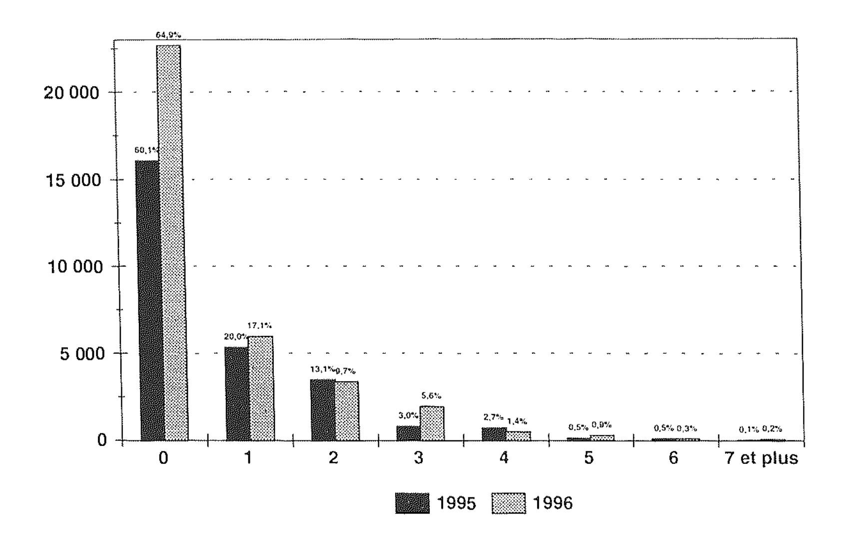
Comme dans l'IPC, les séries du panel ne sont pas systématiquement observables les douze mois d'une année. Le graphique 5 donne par exemple la répartition, en fonction du nombre de mois de présence en 1996, de l'échantillon constitué des séries présentes en décembre 1995.

Graphique 5 : Répartition, en 1996, des séries du poste « café » selon leur durée de présence dans le panel



Il apparaît que 65 % de ces séries ont été observées sans discontinuer (60 % en 1995), 7,6 % ont pu être suivies pendant 11 mois, etc ... La présence d'une série est plus ou moins continue. Le graphique 6 montre ainsi que 17,1 % des séries (présentes en décembre 1995) ont quitté l'échantillon pour ne plus y revenir (du moins en 1996), 9,7 % des séries ont quitté une fois l'échantillon puis l'ont réintégré, 5,6 % l'ont quitté, puis réintégré et quitté à nouveau mais définitivement, etc ...

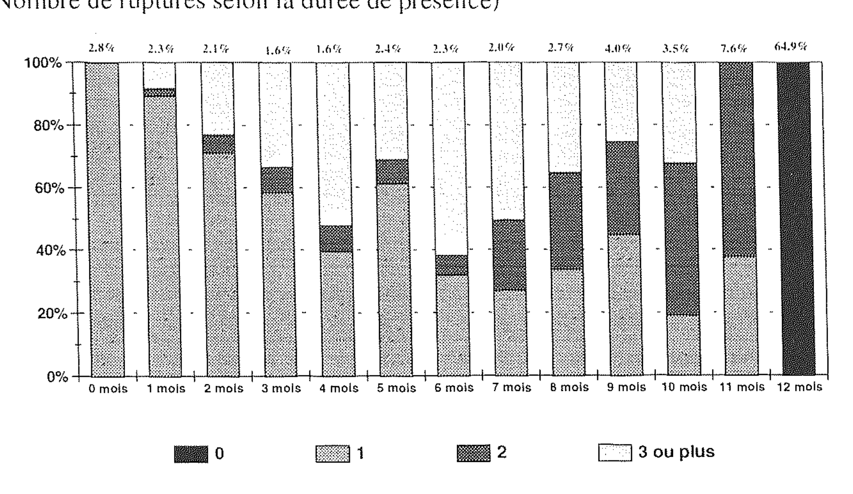
Graphique 6 : Répartition des séries selon leur nombre de ruptures



Le graphique 7 précise les deux précédents. Il indique, par exemple parmi les séries dont la durée de présence a été de 10 mois en 1996, qu'environ 20 % ont été observées durant les 10 premiers mois de l'année (elles étaient donc manquantes en novembre et décembre), un peu moins de 50 % ont été manquantes deux mois consécutifs mais étaient présentes en décembre ; les autres (un peu plus de 30 %) ont quitté l'échantillon un mois, l'on réintégré puis l'ont quitté en décembre (où elles n'ont pas été observées).

Graphique 7 : Stabilité des séries dans le panel en 1996

(Nombre de ruptures selon la durée de présence)



Plus généralement, on tire trois enseignements de ce graphique :

- dans la limite d'une demi-année environ, plus la durée de présence d'une série est longue, plus cette présence est irrégulière ;

- la probabilité d'une unique absence temporaire (i.e. deux ruptures) augmente nettement lorsque la présence dans l'échantillon est supérieure à six mois ;

- ce sont les séries dont la durée de présence est d'environ six mois qui sont les plus instables dans l'échantillon (au moins trois ruptures).

3.4. Une analyse de l'instabilité des séries

Il est difficile, à partir du seul panel, de savoir pourquoi une série, auparavant observée, n'est plus observable un mois donné. Il se peut que le point de vente soit fermé ou bien que la chaîne de distribution à laquelle ce point de vente appartient ne souhaite plus diffuser le produit ; la raison peut également se situer du « côté des produits » : le fabricant est, par exemple, en rupture de stock, ou bien cesse la production d'un produit.

On peut toutefois essayer de mesurer la part de chacun de ces deux effets. Considérons l'ensemble S_{t-1} des N_{t-1} séries observables au cours du mois *t-1* (les séries de *référence*). Le mois suivant *t*, la proportion de séries observables par rapport au mois *t-1* est :

$$\delta_t = \frac{1}{N_{t-1}} \sum_{(i,j) \in S_{t-1}} \delta_t^{i,j}$$

où l'indicateur $\delta_t^{i,j}$ prend la valeur 1 si la série (*i,j*) est observable en *t* et 0 sinon (*i* désigne un produit élémentaire et *j* un point de vente). Le *taux d'absence* des séries est donc $\tau_t = 1 - \delta_t$. Puisque δ_t reste proche de 1 (cf. graphique 8), on a l'approximation :

$$\tau_t \approx \delta_t(1 - \delta_t) \;=\; \text{variance empirique de la distribution } \delta_t^{i,j}$$

Elle permet d'expliquer la valeur du taux d'absence τ_t. Il suffit en effet de procéder à l'analyse de la variance de la distribution $\delta_t^{i,j}$.

Soit, pour cela, S_{t-1}^{i} l'ensemble (à N_{t-1}^{i} éléments) des points de vente de référence d'un produit élémentaire i, c'est à dire l'ensemble des lieux d'achats dans lesquels le produit i était disponible le mois précédent ($S_{t-1}^{i} = \{j, (i,j) \in S_{t-1}\}$). Le taux de présence du produit élémentaire i le mois t dans ses points de vente de référence est :

$$\delta_t^i = \frac{1}{N_{t-1}^i} \sum_{j \in S_{t-1}^i} \delta_t^{i,j} .$$

Soit :

$$\pi_{t-1}^i = N_{t-1}^i / N_{t-1}$$

la part des séries de référence relative au produit élémentaire i. On a :

$$\delta_t = \sum_i \pi_{t-1}^i \delta_t^i$$

et la quantité :

$$V_t^{produits} = \sum_i \pi_{t-1}^i (\delta_t^i - \delta_t)^2$$

mesure la dispersion, *entre les produits élémentaires*, de leur taux de présence le mois t dans leurs points de vente de référence. De l'analyse de la variance de la distribution $\delta_t^{i,j}$ il résulte que

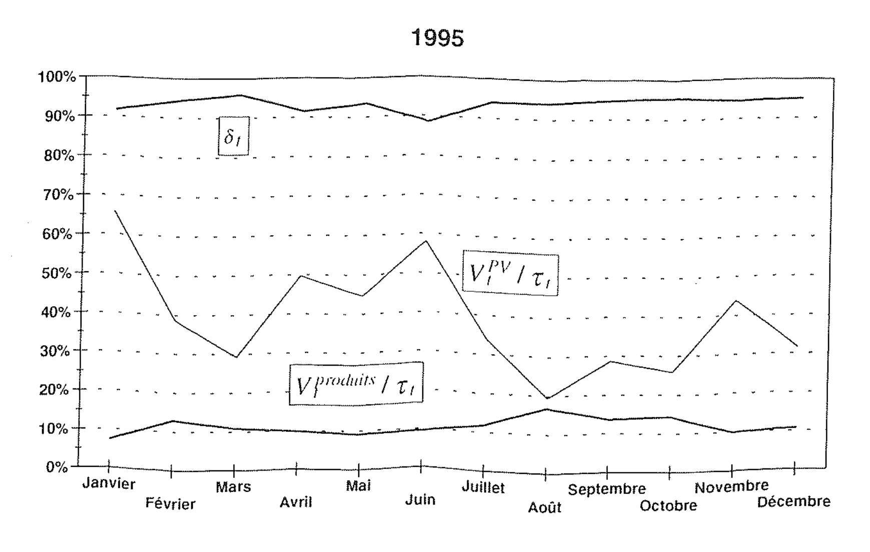
$$V_t^{produits} \leq \tau_t$$

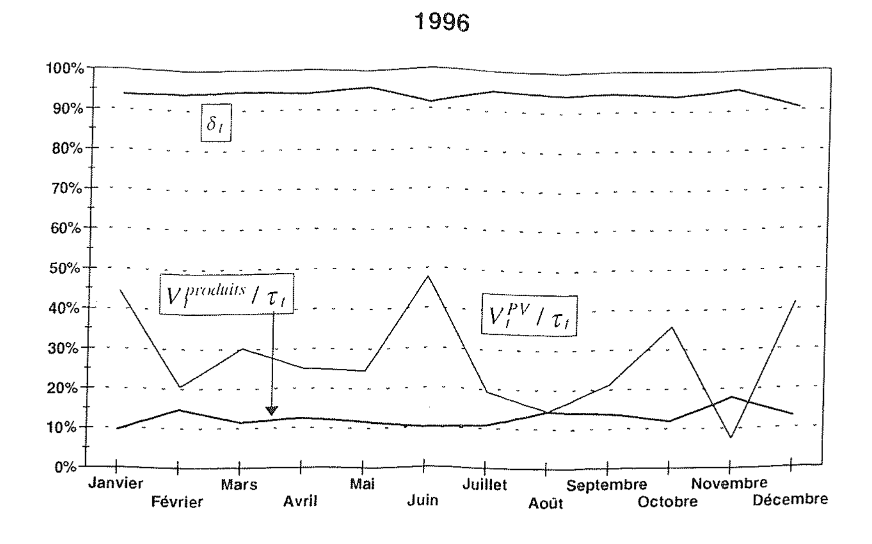
Cette inégalité montre que, mécaniquement, la dispersion $V_t^{produits}$ joue à la hausse du taux d'absence τ_t. Une valeur élevée du ratio $V_t^{produits} / \tau_t$ signifie que la dispersion, *entre les produits élémentaires*, de leur taux de présence dans leurs points de vente de référence explique largement le taux d'absence des séries.

En inversant les rôles des produits et des points de vente, on obtient une mesure de l'effet "point de vente" sur le taux d'observabilité des séries. Cet indicateur

V_t^{PV} / τ_t est d'autant plus grand que la dispersion le mois t, *entre les points de vente*, des taux de distribution des produits qu'ils distribuaient le mois précédent explique largement le taux d'absence des séries.

Graphique 8 : Evolution de δ_t, de $V_t^{produits} / \tau_t$ et V_t^{PV} / τ_t

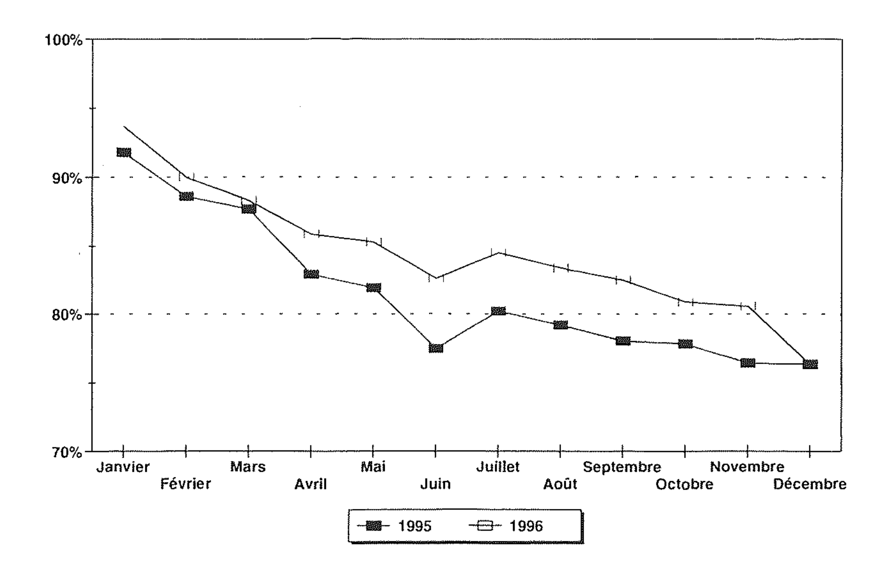




Le graphique 8 montre que l'absence de séries par rapport au mois précédent résulte d'abord de valeur de la dispersion V_t^{PV}. On observe ainsi, sur ce graphique, que τ_t et V_t^{PV} fluctuent en phase (les fluctuations de τ_t sont opposées à celles de δ_t, qui apparaissent sur le graphique) : l'accroissement de la dispersion entre points de vente des taux de distribution des produits élémentaires, en juin et novembre 1995 puis en juin, octobre et décembre 1996, s'accompagne chaque fois d'une hausse du taux d'absence des séries. Un élément d'explication peut être avancé : alors que les fabricants cherchent à écouler leur production par tous les canaux de distribution, les distributeurs semblent se montrer plus sélectifs quant aux produits qu'ils choisissent d'offrir sur le marché.

Si l'on prend maintenant pour séries de référence, l'ensemble "fixe" des séries observables le mois de décembre de l'année précédente (34 930 séries en décembre 1995) plutôt que celles du mois précédent, on constate que le taux de présence de ces séries baisse assez *régulièrement* tout au long de l'année (cf. graphique 9). En effet, d'après le graphique 6, en gros 60 % des séries de référence restent présentes toute l'année, le reste se partageant à peu près également entre les séries qui quittent définitivement l'échantillon après une présence continue, et celles qui ont un comportement plus complexe. La première catégorie de séries stabilise le taux de présence des séries, la seconde catégorie tend à sa baisse *régulière*, quant à la troisième catégorie, elle ajoute un bruit à cette tendance.

Graphique 9 : Evolution annuelle de δ_t



4. La gestion des séries manquantes

L'instabilité des séries est un problème classique dans la construction d'indices de prix : on doit suivre l'évolution des prix des séries d'un échantillon choisi à la période de base. Il faut donc mettre en place, d'une façon ou d'une autre, une procédure de remplacement. Dans l'IPC, l'une des techniques consiste à affecter à une série manquante l'évolution moyenne des prix des séries relatives à la même variété. C'est également l'approche retenue par certains auteurs lors d'études utilisant des micro-données (De Haan-Opperdoes (1997a)[15].). Une autre approche consiste à « agréger » les séries. Cette opération concerne ici les points de vente.

4.1. Agrégation des points de vente par formes de vente

Le nombre de points de vente (hypermarchés et supermarchés) est élevé dans le panel. Or, on a vu à la section 3.4. que la multiplicité des points de vente contribue notablement à accroître le nombre des séries manquantes. En outre, comme on va le voir maintenant, la dispersion des prix des séries entre points de vente est faible, ce qui accrédite l'idée que les substitutions sont essentiellement réalisées entre produits.

Soient N_m le nombre total de séries (i,j) observables un mois m donné et

$$\overline{p}_m = \frac{1}{N_m} \sum_{i,j} p_m^{i,j}$$

leur prix moyen (dans le développement qui suit, chaque série (i,j) est comptée un nombre de fois égal à λ_J^m où J est la strate à laquelle j appartient). On peut décomposer la variance empirique du prix des séries :

$$V_m = \frac{1}{N_m} \sum_{i,j} (p_m^{i,j} - \overline{p}_m)^2$$

en deux termes :

$$V_m = V_m^{produits} + v_m^{produits}$$

15. Comme dans l'IPC Néerlandais, De Haan-Opperdoes (1997a) considèrent des micro-indices qui n'intègrent aucune pondération.

Dans cette relation, $V_m^{produits}$ désigne la variance empirique « inter » produits élémentaires :

$$V_m^{produits} = \sum_i \pi_m^i (\overline{p}_m^i - \overline{p}_m)^2$$

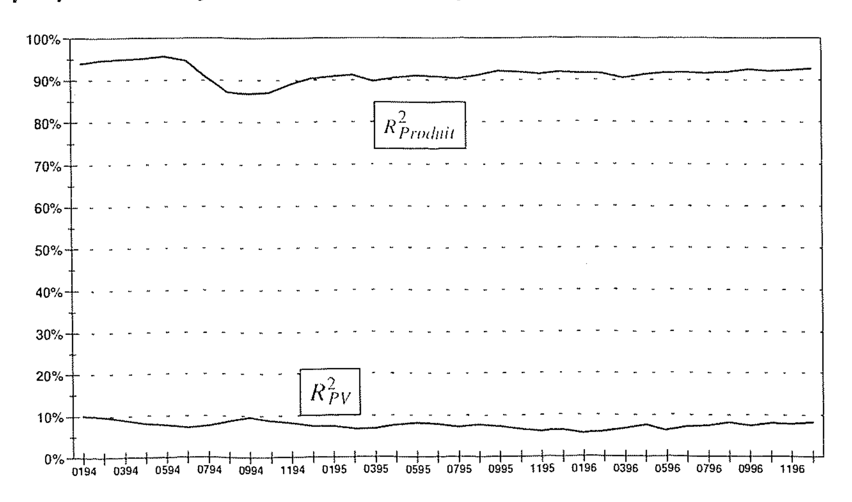
où

$$\overline{p}_m^i = \frac{1}{N_m^i} \sum_j p_m^{i,j}$$

est le prix moyen du produit élémentaire i entre les différents points de vente j où il est vendu (N_m^i est le nombre de points de vente où i est vendu durant le mois m et $\pi_m^i = N_m^i / N_m$). Quant au terme $v_m^{produits}$, il est égal à la moyenne des variances empiriques « intra » des produits élémentaires entre points de vente.

Le graphique 10 indique la part $R^2_{produits} = V_m^{produits} / V_m$ de la variance totale des prix expliquée par la dispersion des prix moyens des produits élémentaires et la part R^2_{PV} expliquée par la dispersion entre points de vente des prix moyens des produits qui y sont vendus.

Graphique 10 : Analyse de la variance des prix des séries



La variance des prix expliquée par les écarts de prix entre points de vente est très faible. On a donc agrégé les points de vente à l'intérieur de chacune des quatre formes de vente (FV). On obtient ainsi 3 127 séries (produits élémentaires, FV). Les prix mensuels moyens de ces séries ont été calculés comme suit :

$$p_m^{i,FV} = \frac{\sum_{j \in FV \cap S^m(J)} \lambda_m^{J(j)} \, ca_m^{i,j}}{\sum_{j \in FV \cap S^m(J)} \lambda_m^{J(j)} \, q_m^{i,j}} \qquad (FV = HM+, HM-, SM+, SM-)$$

où $J(j)$ désigne la strate du point de vente j (cf. Section 2.3.). Grâce à l'utilisation des coefficients d'extrapolation $\lambda_m^{J(j)}$ (cf. Section 2.3.), l'estimation des prix des séries est plus précise (de façon évidente, dans l'expression 2.3. l'utilisation des coefficients d'extrapolation est inutile).

4.2. Elimination des produits nouveaux ou bien retirés de la vente

Afin de pouvoir calculer des indices, on a éliminé les couples (produits élémentaires, FV) qui ne sont pas restés présents pendant 36 mois dans le panel. Il ne restait plus alors que 895 séries (représentant tout de même 92 % du chiffre d'affaires (extrapolé) total des points de vente sur les trois années). Compte tenu de l'agrégation par forme de vente, les séries éliminées correspondaient à des produits nouveaux ou bien retirés, parfois temporairement, du marché. Les produits nouveaux induisent un biais différent du biais de substitution (cf. Turvey (1998)) ; il était donc naturel, dans cette étude, de les écarter.

La répartition des séries entre variétés est donnée dans le tableau 3.

Tableau 3 : Nombre de séries (produit élémentaire, FV) par variété

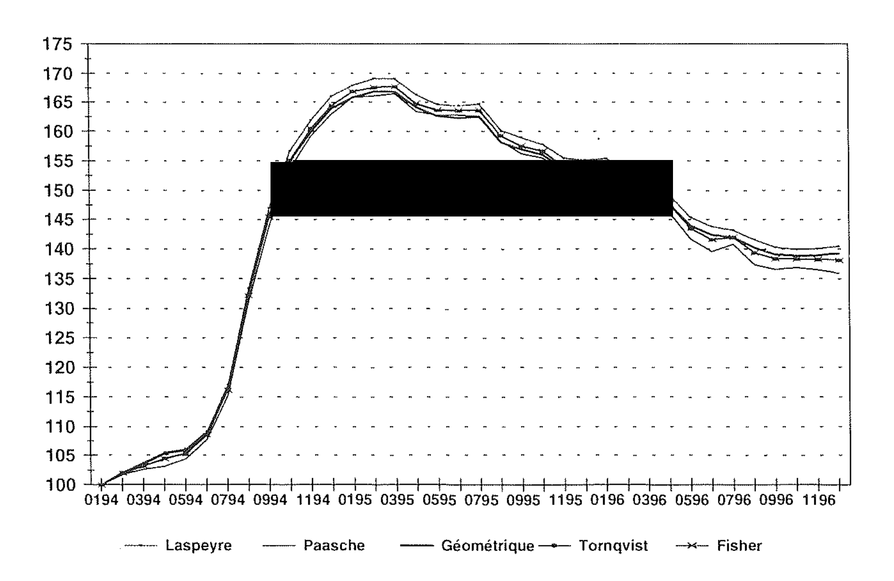
Variétés	nombre de séries(produit élémentaire, FV)
1	153
2	109
3	377
4	135
5	121
Poste	895

5. Calculs d'indices directs et chaînés

5.1. Les indices directs

La définition des indices de Laspeyres, Paasche, Fisher, Törnqvist et Géométrique a été donnée dans l'introduction (dans les formules de définition, *s* parcourt l'ensemble des couples (produit élémentaire, FV) relatifs à une variété ou au poste). Le graphique suivant présente ces différents indices *directs* au niveau du poste :

Graphique 11 : Les indices directs du poste « café »
(Base 100 en Janvier 1994)



Les indices de Laspeyres et de Paasche encadrent les trois autres indices, réputés être proches de l'IUC :

$$P^D < G^D, F^D, T^D < L^D$$

L'indice de Laspeyres sur-pondère en effet les prix qui augmentent le plus alors que l'indice de Paasche sur-pondère ceux qui augmentent le moins. Les indices de Fisher et de Törnqvist sont souvent considérés comme de très bonnes approximations de l'IUC. L'un ou l'autre est pris comme référence pour estimer le biais de substitution, qui mesure l'écart entre l'indice de Laspeyres direct et l'IUC. En décembre 1996, les indices directs de Laspeyres, Törnqvist et Fisher valaient respectivement 140,49, 138,14 et 138,22, soit un biais de 2,35 et 2,27 du Laspeyres par rapport aux deux

autres indices. En pourcentage annuel, le biais est de 0,56 % par rapport à l'indice de Törnqvist et de 0,54 % par rapport à l'indice de Fisher. Mensuellement, le biais de substitution est de 0, 047 % et 0,045 %. A l'intérieur du poste, le biais est assez variable (Tableau 4).

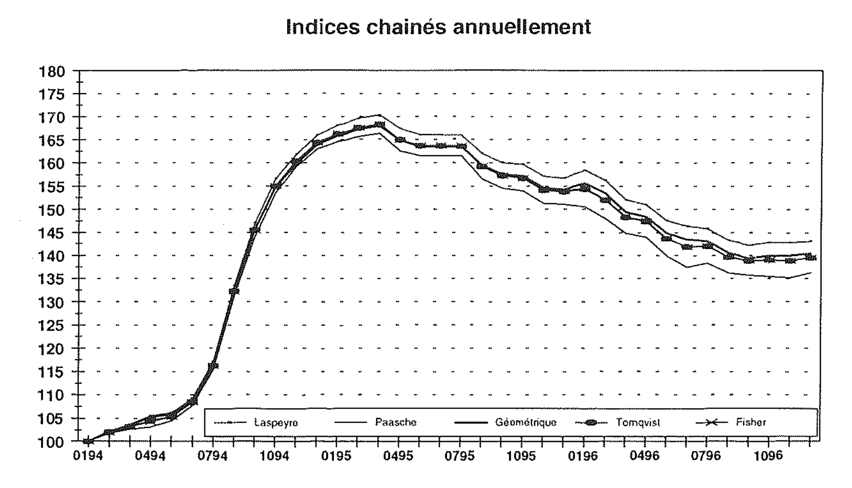
Tableau 4 : Biais de l'indice de Laspeyres direct, pour le poste et les variétés

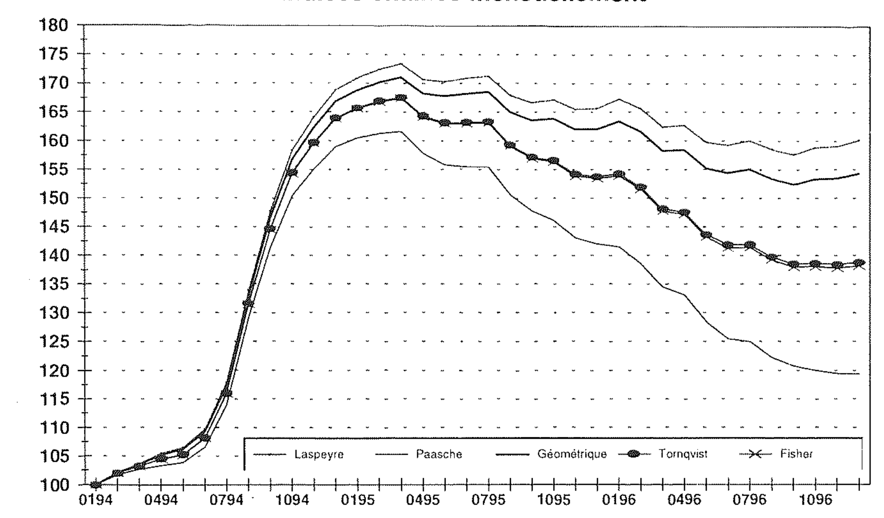
	par rapport à l'indice de Fisher		par rapport à l'indice de Törnqvist	
	par mois (%)	par an (%)	par mois (%)	par an (%)
Variété 1	0,040	0,48	0,04	0,48
Variété 2	0,062	0,75	0,064	0,78
Variété 3	0,036	0,43	0,038	0,45
Variété 4	0,005	0,06	0,003	0,04
Variété 5	0,024	0,28	0,023	0,28
Poste	0,045	0,54	0,047	0,56

5.2. Les indices chaînés

Les faisceaux d'indices directs et chaînés annuellement (comme l'IPC) sont proches. Par contre, ce faisceau s'ouvre largement avec le chaînage mensuel :

Graphique 12 : Les indices du poste chaînés annuellement et mensuellement
(Base 100 en Janvier 1994)

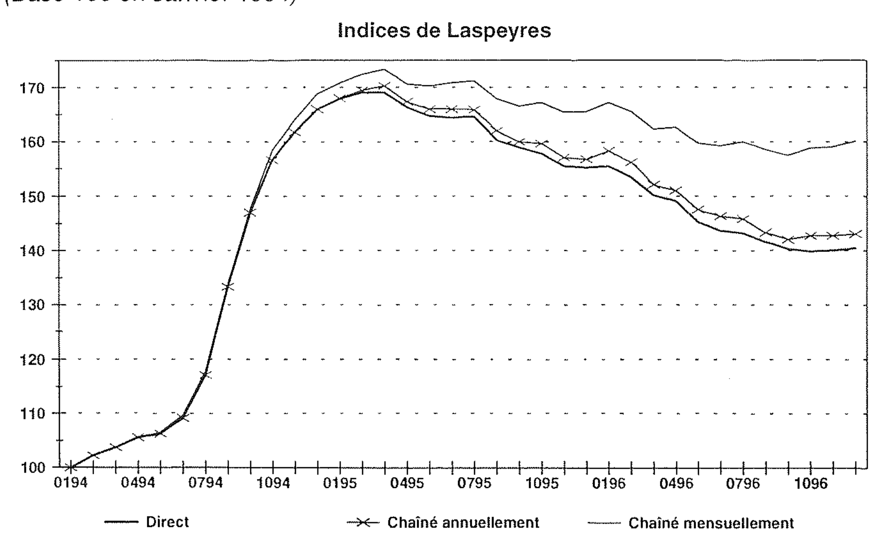


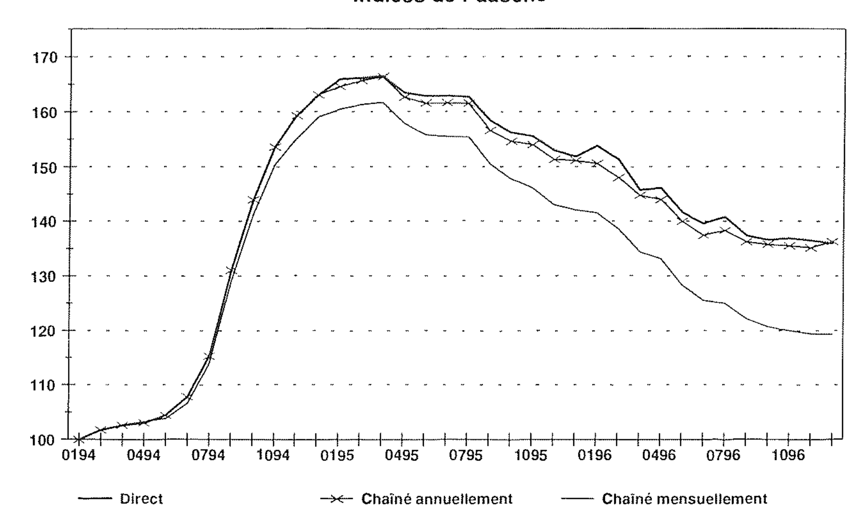


Le constat est le même au niveau des variétés, à l'exception de la variété 2 et, dans une moindre mesure, de la variété 4.

L'indice de Laspeyres chaîné est supérieur à l'indice de Laspeyres direct (l'indice de Paasche chaîné est inférieur au direct). Cette dérive devient très importante avec un chaînage mensuel (graphique 13).

Graphique 13 : Les indices de Laspeyres et de Paasche, directs et chaînés, pour le poste
(Base 100 en Janvier 1994)





On observe, à plusieurs reprises, des évolutions contraires des indices direct et chaîné mensuellement. Ce phénomène est assez régulier (tous les 3 à 5 mois), il ne se produit que dans la phase de baisse de l'indice (après mars 1995) ; il se caractérise par une *hausse* (baisse) de l'indice de Laspeyres (Paasche) chaîné *mensuellement* (et parfois annuellement) quand l'indice *direct* enregistre une *baisse* (hausse).

Comme le biais de substitution, la dérive du chaînage est très variable d'une variété à l'autre, notamment avec le chaînage mensuel (Tableau 5). Pour ce type de chaînage, la variété 2 (café moulu décaféiné) est atypique : après une légère dérive passagère, l'indice chaîné mensuellement reste inférieur aux indices directs et chaînés annuellement. Ce phénomène est étudié en détail à la section 6.

Tableau 5 : Dérive du chaînage de l'indice de Laspeyres, pour le poste et les variétés

	Dérive du chaînage annuel		Dérive du chaînage mensuel	
	par mois (%)	par an (%)	par mois (%)	par an (%)
Variété 1	0,113	1,369	0,677	8,433
Variété 2	0,028	0,335	-0,029	-0,353
Variété 3	0,033	0,402	0,331	3,795
Variété 4	0,037	0,439	0,090	1,089
Variété 5	0,020	0,241	0,213	2,583
Poste	0,051	0,616	0,364	4,455

6. Analyse de la dérive du chaînage

Afin d'analyser le phénomène de dérive de l'indice de Laspeyres chaîné, B. Szulc (1983) a utilisé la relation suivante :

$$L^{C}_{T/0} \;/\; L^{D}_{T/0} = \prod_{t=1}^{T}(1+\rho_t) \text{ avec } \rho_t = \mathrm{cv}(x_t)\mathrm{cv}(y_t).\, r^{t}_{xy}$$

où

$$x_t = (x^s_t)_s \text{ et } y_t = (y^s_t)_s \qquad \text{avec } x^s_t = \frac{p^s_t}{p^s_{t-1}} \text{ et } y^s_t = \frac{q^s_{t-1}}{q^s_0}$$

(s désigne une série) et

$$r_{xy} = (r^t_{xy})_t \text{, avec } r^t_{xy} = \frac{\sum_s \tilde{c}^s_t (x^s_t - \bar{x}_t)(y^s_t - \bar{y}_t)}{\sigma(x_t)\sigma(y_t)}$$

$$\mathrm{cv}(x_t) = \frac{\sigma(x_t)}{\bar{x}_t} \text{ ; } \mathrm{cv}(y_t) = \frac{\sigma(y_t)}{\bar{y}_t}$$

en posant

$$\bar{x}_t = \sum_s \tilde{c}^s_t x^s_t \text{ et } \bar{y}_t = \sum_s \tilde{c}^s_t y^s_t$$

$$\sigma(x_t) = \sqrt{\sum_s \tilde{c}^s_t (x^s_t - \bar{x}_t)^2} \text{ et } \sigma(y_t) = \sqrt{\sum_s \tilde{c}^s_t (y^s_t - \bar{y}_t)^2}$$

Les pondérations $\tilde{c}^s_t$ sont définies de la façon suivante :

$$c^s_t = \frac{p^s_{t-1} q^s_0}{\sum_s p^s_0 q^s_0}$$

$$\tilde{c}_t^s = \frac{c_t^s}{\sum_s c_t^s}$$

Comme $cv(x_t) > 0$ et $cv(y_t) > 0$, l'ordre entre indices de Laspeyres direct et chaîné dépend du signe de la séquence des corrélations r_{xy}^t entre les variables x_t et y_t, c'est à dire entre la variation des quantités vendues entre la période de base et la période précédente, et la variation de prix depuis la période précédente. Szulc a introduit la variable auxiliaire :

$$z_t = (z_t^s), \text{ où } z_t^s = \frac{p_{t-1}^s}{p_0^s}$$

puis considéré que le coefficient de corrélation r_{xy}^t est généralement de même signe que le produit du coefficient de corrélation r_{yz}^t entre y_t et z_t et du coefficient de corrélation r_{xz}^t entre x_t et z_t. Il fonde cette considération sur l'inégalité :

$$r_{xz}r_{yz} - \sqrt{\left(1-r_{xz}^2\right)\left(1-r_{yz}^2\right)} \leq r_{xy} \leq r_{xz}r_{yz} + \sqrt{\left(1-r_{xz}^2\right)\left(1-r_{yz}^2\right)} \qquad (6.1.)$$

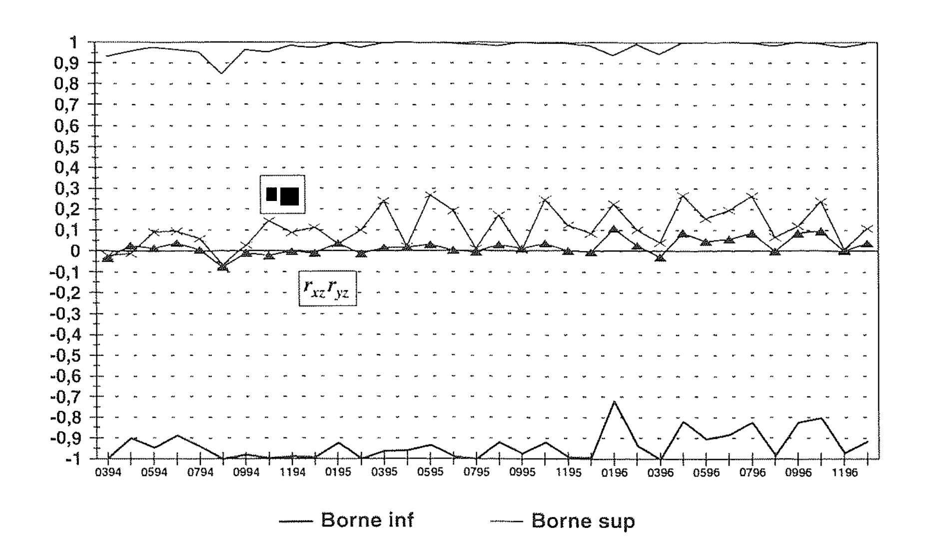
(on omet l'indice t) et l'hypothèse que les carrés des corrélations r_{yz} et r_{xz} sont relativement proches de 1. Puisqu'en général $r_{yz} < 0$ (il est à priori raisonnable de penser que les produits dont les prix ont le plus augmenté, sont ceux dont les quantités vendues ont le plus diminué), un indice de Laspeyres chaîné supérieur à l'indice direct résulterait alors de la persistance ou de la prépondérance d'une situation dans laquelle les produits dont les prix ont le moins augmenté depuis la période de base sont ceux dont les prix augmentent le plus à la période courante.

L'examen de cette analyse à partir du panel AC Nielsen sur le café est intéressant puisque les indices de Laspeyres chaînés, annuellement et plus encore mensuellement, majorent l'indice direct (graphique 13).

Pour le poste, on observe - graphique 14 - une positivité quasi-systématique de r_{xy}. Par contre, selon le même graphique, si l'hypothèse d'évolutions parallèles de r_{xy} et $r_{xz}r_{yz}$ (ou du moins de leurs signes respectifs) est satisfaite en 1996, elle ne s'explique en aucune façon par l'inégalité (6.1.) : dans le cas du café, les racines carrées sont largement plus proches de 1 que de 0. Cette inégalité semble peu

efficace pour fonder l'hypothèse d'une évolution conjointe de r_{xy} et $r_{xz}r_{yz}$: si $r_{yz}r_{xz} = \pm 0{,}7$ alors $\sqrt{\left(1-r_{xz}^2\right)\left(1-r_{yz}^2\right)} = 0{,}5$.

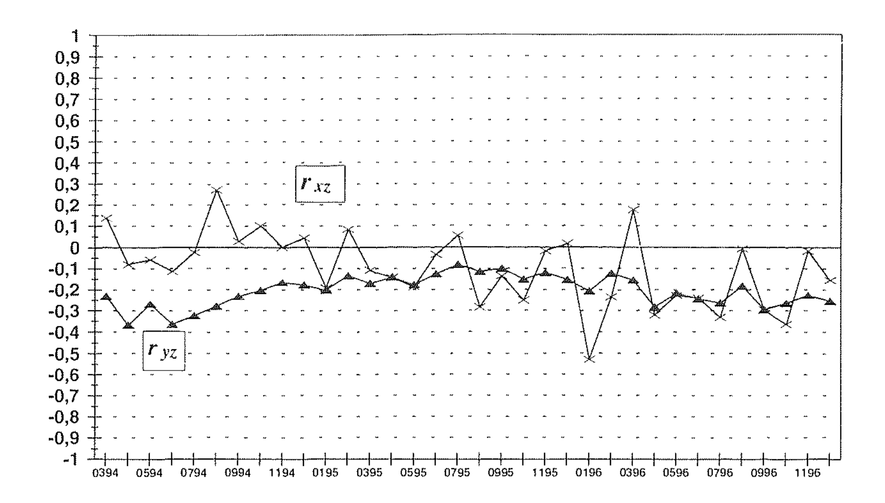
Graphique 14 : Evolutions de r_{xy} et $r_{xz}r_{yz}$ pour le poste (chaînage mensuel)



(Borne inf = $r_{xz}r_{yz} - \sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}$ et Borne sup = $r_{xz}r_{yz} + \sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}$)

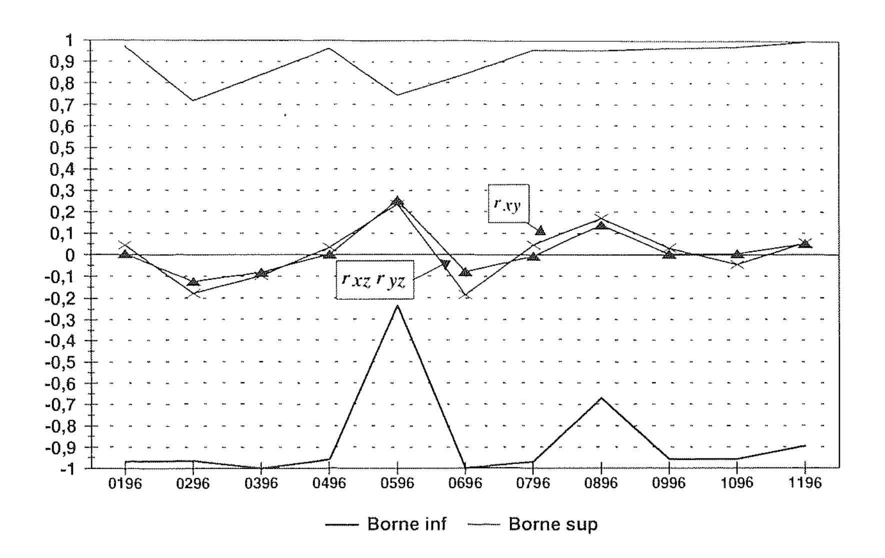
Le cœfficient de corrélation de y et z est toujours négatif (graphique 15), de sorte que la hiérarchie entre indices directs et chaînés dépend de l'évolution de la corrélation entre x et z, du moins sur la période où les évolutions de r_{xy} et $r_{xz}r_{yz}$ sont parallèles. C'est la situation observée en 1996 ; alors r_{xz} est le plus généralement négatif.

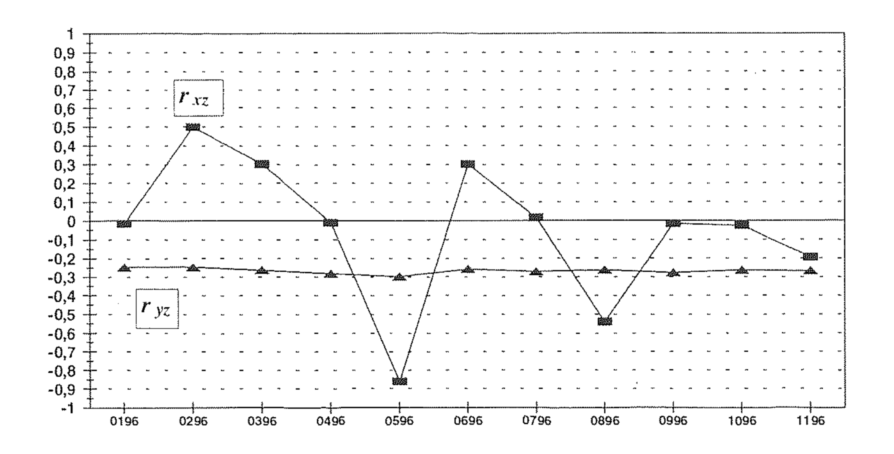
Graphique 15 : Evolutions de r_{xz} et r_{yz} pour le poste (chaînage mensuel)



L'adéquation de l'analyse de Szulc aux données relatives au café est encore plus frappante pour certaines variétés (donc plus mauvaise pour d'autres). A cet égard, la variété 2 est tout à fait remarquable. Cette variété est constituée du café moulu décaféiné. Tout au long de l'année 1996, les coéfficients r_{xy} et $r_{xz}r_{yz}$ restent pratiquement égaux. Là encore ce n'est pas l'inégalité (6.1.) qui justifie cet état de fait : dans le meilleur des cas, $\sqrt{\left(1-r_{xz}^2\right)\left(1-r_{yz}^2\right)} \approx 0{,}3$. Le coéfficient r_{yz} reste stable aux environs de -0,3 (graphique 16).

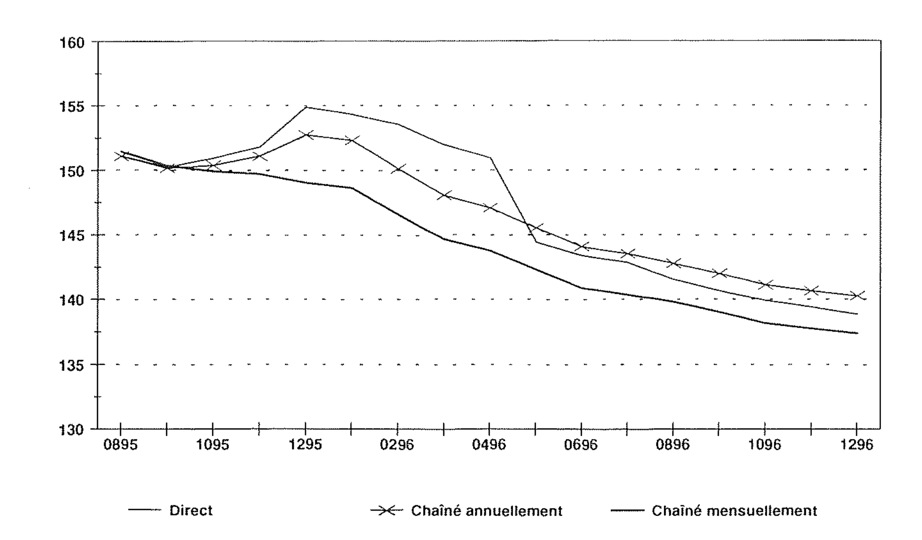
Graphique 16 : Evolutions de r_{xy}, r_{yz}, r_{xz}, et $r_{xz}r_{yz}$ pour la variété 2 (chaînage mensuel)





Les évolutions relatives des indices directs et chaînés sont donc complètement déterminées par celle de la corrélation r_{xz}. En mai 1996, celle-ci est voisine de -1. Si l'indice de Laspeyres direct est supérieur à l'indice chaîné mensuellement, il s'en rapproche brutalement à cette date (graphique 17). Il passe d'ailleurs à ce moment là en dessous de l'indice chaîné annuellement.

Graphique 17 : Les indices de Laspeyres de la variété 2
(base 100 en janvier 1994)



Une mise en garde s'impose : les corrélations calculées dépendent des coefficients c_t^s. Ces coefficients :

$$c_t^s = \frac{p_{t-1}^s q_0^s}{\sum_s p_0^s q_0^s}$$

sont égaux aux parts respectives des séries dans la dépense *initale* des consommateurs, *actualisées par les prix* :

$$c_t^s = \frac{p_0^s q_0^s}{\sum_s p_0^s q_0^s} \frac{p_{t-1}^s}{p_0^s}$$

Elles diffèrent des pondérations *courantes* :

$$\frac{p_{t-1}^s q_{t-1}^s}{\sum_s p_{t-1}^s q_{t-1}^s}$$

Il ne faut pas perdre de vue ce fait dans l'interprétation des corrélations. Ainsi, pour la variété 2, la corrélation r_{vz}, voisine de -1 en mai 1996, est proche de 0 (-0,1) lorsqu'on la recalcule avec les pondérations courantes.

7. Conclusion

L'étude comparative des indices de Laspeyres directs et chaînés a été menée par de nombreux auteurs.

D'abord sur "macro-données", c'est à dire des données utilisées dans le calcul des IPC : prix au mieux relevés mensuellement, pondérations actualisées au mieux annuellement, mais champ étendu à la plus grande partie de la consommation des ménages. Les résultats obtenus sont divers. Les uns montrent que le chaînage a permis de réduire le biais de substitution. Il en est ainsi de Diewert (1978) et Manser-McDonald (1988). Diewert utilise des données en prix et quantités relatives aux dépenses de consommation au Canada sur la période 1947-1971 ; il chaîne tous les cinq ans. Le chaînage élimine la quasi-totalité du biais de substitution évalué avec les indices de Törnqvist ou de Fisher. Manser et McDonald (1988), qui suivent l'approche non paramétrique du calcul de l'IUC (cf. Introduction), obtiennent des résultats analogues pour les Etats-Unis sur la période 1959-1985 : l'écart entre l'indice direct et l'indice chaîné représente près des trois quarts du biais de substitution. Au contraire, les résultats obtenus par Aizcorbe-Jackman (1993), qui portent sur une période plus récente (1982-1991), montrent une légère dérive du chaînage.

L'analyse comparative des indices de Laspeyres directs et chaînés n'a été effectuée sur "micro-données" que récemment, par J. de Haan-E. Opperdoes (1997b), M. Reinsdorf (1995) et J. Dalen (1997) notamment[16]. Les résultats obtenus mettent *tous* en évidence une *importante* dérive du chaînage *mensuel* : loin de réduire le biais de substitution, il l'accentue fortement. Les études de de Haan-Opperdoes et de Reinsdorf portent sur des données scannées relatives au café ; l'étude de Dalen sur les matières grasses, les lessives, les céréales pour le petit déjeuner et le poisson congelé.

L'utilisation, dans la présente étude, du cadre d'analyse de B. Szulc montre que le phénomène de "rebond" des prix autour de leur tendance, c'est à dire leurs fluctuations décalées avec une fréquence élevée, est à l'origine de la dérive du chaînage mensuel. Ceci suggère l'idée d'une valeur optimale de la durée séparant les chaînages successifs. Une durée d'une ou quelques années serait proche de cet optimum : suffisamment longue pour échapper au phénomène de rebond des prix, elle assure une bonne prise en compte des substitutions opérées par les consommateurs.

---

16. Aucun de ces auteurs n'a testé l'analyse de Szulc sur le rebond des prix.

Bibliographie

S.N. Afriat (1967), *The construction of a utility function from expenditure data*, International Economic Review, 8, 125-1333.

A.M. Aizcorbe and P.C. Jackman (1993), *The commodity substitution effect in CPI data, 1982-91*, Monthely Labor Review.

B.M. Balk (1998), *On the use of unit value indices as consumer price subindices*, Departement of statistical methods, Statistics Netherlands, Voorburg.

J. Dalen (1997), *Experiments with swedish scanner data*, International Working Group on Price Indices, Statistics Sweden.

W.E. Diewert (1973), Afnat and the Revealed Preference Theory, Review of Economic Studies 40, 419-426.

W.E. Diewert (1976), *Exact and Superlative Index Numbers,* Journal of Econometrics, vol. 4, n°2.

W.E. Diewert (1978), *Superlative Index Numbers and Consistency in Aggregation*, Econometrica, Vol. 46, n°4.

W.J. Hawkes (1995), *Reconciliation of fixed-weight price index trends with corresponding trends in average prices for quasi-homogeneous goods using scanning data*, Nielsen USA, Working paper.

J. de Haan - E. Opperdoes (1997a), *Estimation of the coffee price index using scanning data : simulation of official practices,* Statistics Netherlands.

J. de Haan - E. Opperdoes (1997b), *Estimation of the coffee price index using scanning data : the choice of the micro index,* Statistics Netherlands.

C.R. Hulten (1973), *Divisia index numbers*, Econometrica 41, n°6, 1017-25.

F. Lequiller (1997), *L'indice des prix à la consommation surestime-t-il l'inflation ?*, Economie et Statistique, n°303.

M.E. Manser and R.J MacDonald (1988), *An analysis of substitution biais in measuring inflation, 1959-1985*, Econometrica, Vol. 56, n° 4, 909-930.

M.B. Reinsdorf (1995), *Constructing basic component indexes for the US CPI from scanner data : a test using data on coffee,* U.S. Bureau of Labor Statistics.

M.B. Reinsdorf (1998), *Divisia Indexes and the Representative Consumer Problem*, Working Paper.

A. Saglio (1995), *Changement de tissu commercial et mesure de l'évolution des prix*, Economie et Statistiques, n° 285-286, pp 9-33.

B.J. Szulc (1983), *Enchaînement des indices de prix*, in *La mesure du niveau des prix*, actes du colloque tenu sous l'égide de Statistic Canada, W.E. Diewert et C. Marquette eds, Canada.

R. Turvey (1998), *New outlets and new products*, Proceedings of the Third Meeting of the International Working Group on Prices Indices, B.M. Balk editor, Statistics Netherlands.

L'UTILISATION DE LA VALEUR UNITAIRE COMME INDICE DE PRIX DES SERVICES AUX ENTREPRISES : LE CAS DES TÉLÉCOMMUNICATIONS

Charles Bérubé

Comme nous le savons tous, le domaine des télécommunications est en plein essors dans les pays industrialisés. De plus, l'ouverture des marchés à la concurrence, dans un secteur où autrefois l'état (ou bien une compagnie réglementée par ce dernier) détenait un monopole, entraîne des changements radicaux. Les services offerts sont de plus en plus nombreux. Ils se vendent à des prix qui changent de plus en plus rapidement en même temps que le service lui même évolue. En effet, la concurrence engendre parfois un changement dans le type de service offert sans que nécessairement il y ait une évolution de tarification. Par exemple une compagnie de téléphone pourrait offrir un an d'abonnement à internet si l'abonné reste fidèle. Elle peut aussi offrir un ensemble de service (ligne de téléphone, ligne télécopieur, accès internet, téléphone mobile, etc.) avec des tarifs variables selon l'ensemble choisi. Les tarifs, le contenu ainsi que les conditions afférentes à ces ensembles de services peuvent varier d'un mois à l'autre selon la stratégie du concurrent. On parle même de compagnies de téléphone pouvant offrir, grâce à la fibre optique, les mêmes services que les câblodistributeurs. Il n'est donc pas impossible dans le futur que votre compagnie de téléphone vous offre en plus du téléphone, plusieurs chaîne télé, un service de film à domicile et un accès internet pour un seul prix.

En général les indices de prix calculés par les agences statistiques, que ce soit un indice à la production ou bien à la consommation, se veulent des indices permettant de mesurer l'inflation[1] d'un panier de biens et services. Le but étant de calculer des mouvements de prix pures, ont choisi généralement un panier de biens et services représentatifs à une période 0 que l'on réexprime consécutivement en prix courants (période t). On peut très bien imaginer les difficultés rencontrées pour calculer un mouvement de prix pures lorsque les services sont en constante évolution. Le panier n'est plus vraiment fixe et devient de moins en moins comparable d'une période à l'autre.

---

1. La polémique entre l'utilisation d'une mesure d'inflation ou bien celle d'une mesure du coût de la vie dans le cas d'un indice de prix à la consommation dure toujours. Par contre l'indice des prix à la consommation de la plupart des agences statistiques officielles reste une mesure d'inflation et non celle du coût de la vie. Le but de cette communication n'est pas de débattre ce sujet mais porte plutôt sur les indices de prix à la production.

Le but de ce papier sera de discuter l'utilisation d'une valeur unitaire afin de calculer un indice de prix à la production pour les services de télécommunication. Nous tenterons d'expliquer pourquoi cette approche de valeur unitaire est une alternative acceptable mais moins rigoureuse que l'approche du panier fixe. Nous discuterons aussi des avantages et désavantages d'un indice à valeur unitaire, les techniques qui peuvent être prise pour calculer l'indice, et l'utilisation des technologie ou bien des coûts de production pour regrouper les services de télécommunication en catégories homogènes.

L'utilité première d'un indice de prix sur les services de télécommunications, est de mesurer les mouvements de prix dans le temps afin de générer un déflateur pour les comptables nationaux. L'approche proposée repose sur la théorie du producteur[2] qui supposent que les producteurs de services de télécommunications maximisent leur profit étant donné les contraintes technologiques (i.e. contraintes de coût).

Expériences passées

Un projet pilote a été entrepris à Statistique Canada en collaboration avec la compagnie de téléphone Bell Canada. Ce dernier voulait que la division des prix à Statistiques Canada incorpore les nouveaux rabais (sous forme de plan d'économie) à l'indice des prix à la consommation. La manoeuvre quoique justifiée consistait à pouvoir révéler au conseil de radio et télédiffusion3 une nette tendance à la baisse du prix des interurbains afin de justifier une hausse de prix des communications locales. A la division des prix, nous avons donc profité de l'occasion afin de sensibiliser Bell Canada sur la nécessité d'un indice de prix à la production. Le projet initial consistait à calculer un indice de Laspeyres en chaîne à l'aide d'un panier de services où les quantités restaient fixes. Cette méthode permettrait donc de calculer des mouvements de prix purs.

En général Bell offre à ses clients différents plans qui peuvent être composés de plusieurs services. Ces services sont aussi composés d'interurbains et de services locaux. Les interurbains sont la plupart du temps divisés en quatre soit: Les appels à l'intérieur du territoire de la compagnie, ceux à l'intérieur du territoire des autres compagnies canadiennes, les appels aux USA et enfin les appels outre-mer. Cette complexité rend très difficile l'observation du mouvement de prix d'un service en particulier, à moins de pouvoir observer les factures de clients. Les compagnies de

2. La théorie du producteur et la théorie du consommateur (qui est à la base de l'indice des prix à la consommation) peuvent parfois nous entraîner à des conclusions différentes à moins que certaines conditions assez restrictives tiennent [voir E.Diewert (1983), page 1049 dans "Price level mesurment"]. Ces différences peuvent justifier l'emploi d'une approche différente de celle de l'indice des prix à la consommation.

3. Organisme gouvernemental qui réglemente le marché des télécommunications.

téléphone possèdent les informations nécessaires à la construction d'un indice permettant de calculer un mouvement de prix pur. La possibilité de pouvoir suivre les mouvements de prix à partir de l'univers ou même d'un échantillon de client a été proposée aux compagnies de téléphone sans succès. Même si les données nécessaires à la construction d'un panier fixe existe déjà, ces derniers ont refusé la méthode proposée en invoquant le coût en ressources humaines et matérielles afin de pouvoir extraire, programmer, stocker et maintenir ces données.

Pourquoi une valeur unitaire ?

Même si les compagnie de téléphone ont bien accueilli l'idée d'un indice de prix sur les services aux entreprises, il faut dire que leur définition de prix était différente de la nôtre. Pour la division des prix, le prix est la valeur sur le marché de la plus petite unité de service (ou d'un bien). A l'aide de ce concept, on peut s'attendre à une relation non ambiguë entre le vecteur des prix unitaire et le vecteur des quantités étant donné les conditions de marché. Toutefois, dès le début des négociations, les compagnies de télécom semblaient insister pour définir le prix comme étant une valeur unitaire des services. Ils le définissait comme un ratio du revenu total généré par les ventes des services de télécom pour un plan donné[4] sur la quantité totale d'output transigée:

$$P_i^u = \frac{\sum_i P_i Q_i}{\sum_i Q_i} \quad (1)$$

où P_i^u est la valeur unitaire (exemple: le revenu moyen par minute pour le service). Les compagnies de télécom utilisent P_i^u comme un outil marketing mais pour un statisticien, l'idée d'utiliser une valeur unitaire, spécialement pour la construction d'un indice de prix, n'est pas très rassurante (i.e. la valeur unitaire peut varier même si le prix ne change pas). L'unité de quantité Q_i^u peut être des minutes, une fréquence d'usage, le nombre de ligne, etc. Par contre, les différentes mesures d'unité ne peuvent pas être combinées.

Comme nous l'avons mentionné plus haut l'idée de pouvoir associer un service particulier à un tarif correspondant ne semble pas réalisable du point de vue des compagnies de télécom. Au moins trois raisons spécifiques ont été donné. Premièrement, le prix de transaction qui est chargé à l'entreprise dépend du type et de l'étendue des plans d'escompte offert dans un marché compétitif. Très souvent les

4. Par plan on entend un ensemble de services, un "package deal".

plans offerts sont liés à des escomptes qui varient avec la journée, l'heure dans la journée, le volume des appels, le schéma de consommation et les accords signés avec les clients (i.e. contrats). Ces escomptes font partie du prix de transactions et ne peuvent pas être ignorés. Il semble presque impossible[5] de séparer les escomptes des tarifs pour chaque type de service car, très souvent, l'escompte n'est pas associé à un service mais à un ensemble de services. Deuxièmement, on peut noter la facilité avec laquelle les compagnies peuvent "jouer" sur les modalités des plans d'escompte. Il est peu probable que ces dernières se présentent devant la commission de régulation pour faire modifier leur tarifs. Les tarifs deviennent donc des valeurs fictives qui ne reflète pas les vrais conditions du marché. Ils représentent donc mieux les coûts de production qui sont fonction de la distance pour les interurbains, de l'heure dans la journée ainsi que du type de service. Troisièmement, les efforts demandés pour désagréger systématiquement une combinaison de prix, de distance et de temps sont trop grands.

En fin de compte, nous n'avons pu que constater l'échec. La construction d'un indice de prix des services de télécom aux entreprises calculant des mouvements de prix purs semble difficilement réalisable si on veut vraiment tenir compte de l'état du marché actuel.

Dans le contexte d'un indice de prix à la consommation la solution alternative qui consiste à calculer des valeurs unitaires ne nous semble pas acceptable du fait même de l'utilisation de l'indice. Par contre cette solution que l'on pourrait qualifier de "second best" pourrait s'avérer une alternative utile et peu coûteuse pour l'indice des services aux entreprises. La valeur unitaire en tant que concept comporte certains avantages mais aussi certaines limites. Les limites potentielles doivent toutefois être jugées en regard des bénéfices obtenus. Le bénéfice le plus évident serait d'obtenir un déflateur plus fiable que celui qui est actuellement utilisé[6] par les comptables nationaux canadiens. Dans un certain sens, avoir conscience des limites de la valeur unitaire en tant qu'indice nous force à trouver des solutions qui éventuellement nous permettrons d'atteindre un niveau de confiance acceptable quant à l'utilité et à la pertinence de l'indice. Ces solutions sont discutées dans la section méthodologie.

De plus, il est important de souligner ces autres avantages: En premier lieu, la valeur unitaire tient compte des divers types d'escompte. En second lieu, elle peut être regardée comme un prix hybride et utilisée pour construire un indice de prix. Enfin, lorsque les transactions sont assez détaillées, la valeur unitaire fournit un prix moyen de transactions moyennes pour un groupe donné de services.

---

5. Lors des négociations pour l'indice des prix à la consommation, nous avons pu constater que les compagnies de télécom pouvaient générer, par type de service, des revenus agrégés et des minutes de conversation avant et après escompte.
6. La division Input-Output utilise une combinaison du nombre de ligne et du nombre d'appels interurbains pour construire des indices de volume.

Ce dernier point de vue semble être partagé dans la littérature existante. Par exemple, Erwin Diewert (1995) dans un article intitulé "Axiomatic and economic Approaches to Elementary Prices Indexes" commente l'utilisation de la valeur unitaire (lorsque le niveau de désagrégation est suffisant) de la façon suivante: "It should be noted that a unit value for the commodity provides a more accurate summary of an average transaction price than an isolated price quotation" [p.23][7]. Bert Balk (1995)[8] va encore plus loin avec ses six axiomes. Ces axiomes décrivent les caractéristiques mathématiques qui permettent de juger la valeur unitaire en tant qu'indice :

Axiomes	Conclusions	Passe le Test
Homogénéité de degré 0	La valeur unitaire n'est pas dépendante des valeurs absolues des prix et quantités mais plutôt des changements relatifs.	Oui
Homogénéité linéaire	La valeur unitaire change de la même façon lorsque les prix changent dans une proportion de 1.	Oui
Monotonicité	Si les prix augmentent, l'indice augmente.	Oui
Identité	L'indice peut être différent de 1 même si les prix sont identiques d'une période à l'autre.	Non
Proportionnalité	Doubler les prix ne font pas nécessairement doubler l'indice.	Non
Invariance dimensionnelle	Changer l'unité de mesure peut faire varier l'indice.	Non

Méthodologie possible

Deux questions semblent pertinentes afin de définir le cadre méthodologique. Tout d'abord, étant donné les informations qui pourraient être mises à notre disposition par les compagnies de téléphone, comment identifier et regrouper les services de télécom les plus représentatifs de l'industrie ? Et ensuite, quelles procédures pouvons-nous suivre afin de minimiser les biais inhérents à un indice basé sur la valeur unitaire.

Le biais mentionné découlant de l'axiome 4 de Balk introduit des "impuretés" dans l'indice qui peuvent être non négligeables. Une façon de minimiser ces "impuretés" consiste à stratifier les services de télécom en groupes assez homogènes. Mais comment peut-on différencier un groupe homogène de services d'un groupe non homogène ? La stratégie préconisée consiste à regarder le processus de production

7. W.E.Diewert, "Axiomatic and Economic Approaches to Elementary Prices Indexes", Economics department, University of British Columbia, 1995. A paraitre dans le "Journal of Economic Literature" numero de classification: C43, C81,E31, O47.

8. Bert Balk travaille au Centraal Bureau voor de Statistiek à Voorburg au Pays-Bas.

afin d'obtenir une idée de la façon dont les inputs sont combinés pour produire différents services de télécom. La proportion d'inputs utilisée peut varier d'un type de service produit à l'autre mais les producteurs choisissent la proportion optimale. Entre autres, la production de services de télécom implique des fonctions de réseaux, de distributions, de type de transmission (données, voix, images). Les services de télécom peuvent donc être différentiés sur la base d'une homogénéité technologique. Il faut faire attention car les mêmes services peuvent être présentés aux entreprises de différentes manières leur laissant croire qu'il s'agit de services différents. Autrement dit un service peut être offert à différents prix sur différents marchés créant ainsi l'impression qu'il existe plusieurs services différents.

Si on prend le point de vue du producteur, le test concluant devrait être de déterminer comment ces services sont produits. Si les services sont produits à l'aide des mêmes ressources mais sont différenciés par une stratégie de marketing, ils devraient être regroupés quelque soit l'acheteur de ces services. Cette question qui est essentiellement technologique, peut être résolue en examinant la structure des coûts de production. Le critère d'homogénéité technologique a au moins trois avantages : (a) puisqu'on construit un indice à la production, cette approche fournie un cadre de travail conceptuel afin de regrouper les services à un niveau de désagrégation assez fin, (b) l'idée pourrait servir à évaluer les changements de qualité des prix unitaires et (c) tend à réduire l'effet d'hétérogénéité sur la valeur unitaire à l'intérieur des strates(i.e. groupes de services).

Par conséquent il serait plus sage dans un premier temps de construire un indice basé sur une technologie connue où le service offert est bien implanté. Par exemple le service des interurbains sans faire de référence à la technologie des portables. On suppose que les entreprises peuvent avoir accès à plusieurs plans d'économie leur permettant d'effectuer des interurbains. En général, ces interurbains seront chargés selon des critères de zone ou de distance. Ces critères de zone ou de distance peuvent être considérés comme étant le niveau le plus désagrégé pour lequel il est possible d'obtenir de l'information (revenus et quantités).

Un indice basé sur la valeur unitaire pour un plan d'économie impliquerait donc

(a) la sélection des services représentatifs au niveau de désagrégation le plus fin (i.e. Par exemple, inter-région, inter-Europe, International),

(b) le calcul des valeurs unitaires pour chaque service représentatif,

(c) le calcul d'une moyenne géométrique de ces valeurs et

(d) la détermination d'un panier composé de plusieurs plan d'économie.

On peut formuler de manière plus formelle comme ceci :

$$I = \frac{P_t^{MG}}{P_0^{MG}} * \left[\frac{P_0^S Q_0^S}{\sum_S P_0^S Q_0^S} \right] \quad (2)$$

où I est l'indice de base pour un plan d'économie donné S.

$\frac{P_t^{MG}}{P_0^{MG}}$ est la valeur unitaire relative (du plan d'économie S) entre la période de base 0 et la période de référence t calculée avec la moyenne géométrique suivante:

$$P^{MG} = \sqrt[N]{\prod_i^N P_i^U} \quad (3)$$

P_i^U est la valeur unitaire d'un service de base i (ex appels inter Europe) pour un plan d'économie donné S. Cette valeur unitaire se calcule à l'aide de l'équation 1. Finalement chaque valeur unitaire relative est multipliée par sa pondération qui est déterminée à la période de base par l'importance relative du plan d'économie calculée par $\frac{P_0^S Q_0^S}{\sum_S P_0^S Q_0^S}$.

Conclusion

Les biais révélés par les axiomes 4 et 5 ne sont certainement pas négligeables mais pourraient être acceptables. L'impact des biais peut être minimisé si l'on regroupe les services en des catégories homogènes à un niveau de désagrégation le plus fin possible. Si nécessaire, les valeurs unitaires peuvent être combinées à l'aide d'une moyenne géométrique pour un sous-groupe de services faisant partie d'un même plan d'escompte. Par contre, minimiser le biais généré par le non respect de l'axiome de proportionnalité semble moins évident. Même si l'indice ne double pas lorsque les prix sont doublés, l'indice retiendra quand même le mouvement de prix dans le temps. Des études empiriques seraient nécessaires afin d'observer les implications pratiques sur l'indice causé par ce biais. Enfin, il semble qu'à priori un indice de ce type devrait faire l'objet d'une utilisation très restreinte et n'aurait probablement pas droit au titre d'indice de prix.

BIAIS DES INDICES DE PRIX À LA CONSOMMATION : OÙ EN EST-ON ?

François Lequiller

1. Le fond du débat[1]

Aiguillonné par plusieurs années de débat sur une possible surestimation de l'inflation aux Etats-Unis, le Sénat américain a appelé une commission d'économistes, présidée par un professeur de Stanford, M.J. Boskin, à rapporter sur cette question. Cette commission lui a remis un rapport final en décembre 1996 qui soutient l'existence d'une surestimation de la hausse des prix par l'IPC des Etats-Unis de 1,1 % par an pour les années postérieures à 1996 et de 1,3 % pour les années antérieures. En d'autres mots, la « commission Boskin » affirme que la « vraie » inflation au stade de la consommation des ménages serait plus faible de 1,1 % par an au chiffre qui sera publié par le BLS [2] dans les années futures. Par exemple, si l'IPC américain augmentait de 3,0 % l'an prochain, il faudrait comprendre, d'après la commission, que la véritable hausse n'aura été que de 1,9 %.

Comme il s'agirait d'une surestimation qui se cumulerait d'une année sur l'autre, elle serait mécaniquement égale à 11,5 % au bout de 10 ans [3]. Or l'IPC est directement utilisé aux Etats-Unis pour indexer les prestations sociales et les tranches de l'impôt sur le revenu. Dans un cas comme dans l'autre, une surestimation conduirait à un creusement intempestif du déficit fédéral. Ainsi, les prestations versées seraient trop élevées et les impôts reçus seraient minorés (du fait d'une hausse trop rapide des tranches de l'impôt sur le revenu). La commission a estimé que la dette publique serait plus élevée d'environ 1000 milliards de dollars en 2008 de ce simple fait. Dans cette période de débats houleux sur le déficit fédéral, elle n'hésitait pas à qualifier la surestimation de l'IPC de « quatrième poste de dépense du budget fédéral après les prestations sociales, les dépenses de santé et la défense ».

Une telle surestimation aurait des conséquences également très importantes sur de nombreuses autres mesures macro-économiques. C'est le cas par exemple du

1. Cette partie de l'article est un simple résumé de Lequiller, 1997.
2. *Bureau of Labor Statistics*, l'homologue américain de l'Insee pour l'IPC.
3. Cette extrapolation repose sur une hypothèse de constance du biais dans le temps qui n'a pas été vraiment explorée. Elle suppose également que le BLS n'apportera pas des corrections à ses méthodes dans le futur.

nombre de ménages en dessous du seuil de pauvreté. Un économiste américain (Baker, 1996) a ainsi calculé que plus de la moitié des ménages américains auraient été classés en 1962 en dessous du seuil de pauvreté actuel si l'on acceptait les conclusions de la commission Boskin. Ce résultat, peu vraisemblable, a contribué à alimenter la critique des conclusions de la commission.

D'où proviendrait cette surestimation de l'indice des prix ? La commission Boskin, reprenant des travaux très fournis, particulièrement aux Etats-Unis, recensait quatre sources de surestimation.

Tableau 1 : USA: surestimation de l'indice des prix à la consommation

Sources de surestimation	Etats-Unis
	Valeur estimée en % annuel pour les années postérieures à 1996
Substitution au niveau agrégé	0,15
Substitution au niveau détaillé	0,25
Nouveaux produits	0,60
Nouveaux circuits de distribution	0,10
Total	**1,10**

Source : rapport Boskin.

Les deux premières sources de surestimation (substitution aux niveaux agrégé et détaillé) proviendraient d'une seule cause: la mauvaise prise en compte par les formules de calcul et les pondérations utilisées par les indices de prix des effets de la substitution que les ménages opèrent entre produits. La troisième source de surestimation (nouveaux produits) proviendrait de la mauvaise évaluation par les statisticiens des améliorations de la qualité des nouveaux produits. La quatrième et dernière source de surestimation proviendrait d'une mauvaise prise en compte des baisses de prix concomitantes aux gains de parts de marché des nouveaux circuits de commercialisation (super, hypermarchés...).

Substitution entre produits : le problème des pondérations et des formules d'indice

Le problème posé aux constructeurs de l'IPC par la substitution des produits provient de ce que les ménages modifient leur panier de consommation *en même temps* que les prix varient. Comme on va le voir, un IPC dont les pondérations

reposeraient sur des informations obsolètes ou sur des formules de calcul ne tenant pas compte des substitutions pourrait avoir tendance à surestimer l'inflation.

Plaçons-nous d'abord dans le cas d'un consommateur *unique* et essayons de calculer *son* indice de prix entre une période de base et la période courante. On définit l'indice de prix pour ce consommateur entre ces deux périodes comme le taux de croissance de sa dépense budgétaire qui lui permet de conserver, avec les prix courants, le même niveau de satisfaction qu'à la période de base. C'est l'idée de la préservation du « pouvoir d'achat ». C'est aussi l'idée de l'indice « idéal », dit à utilité constante (IUC).

Si tous les prix des produits variaient proportionnellement, cet indice de prix serait très facile à calculer. Il suffirait de choisir l'un des produits et d'en observer la variation de prix. On sait cependant que la hausse (ou la baisse) moyenne des prix masque des variations contrastées entre produits. En d'autres termes, les prix relatifs se modifient avec le temps en même temps que le mouvement général de hausse ou de baisse. Pour calculer l'indice de prix de notre consommateur, il faut donc faire entrer en ligne de compte, sinon la totalité des produits qu'il consomme, au moins un échantillon qui en soit représentatif et effectuer une *moyenne* des variations de prix des produits qui le composent. Se pose alors immédiatement la question de la pondération avec laquelle chacun des produits doit rentrer dans cette moyenne. La pondération qui s'impose est bien sûr fondée sur la quantité consommée. *L'indice de prix est donc le résultat d'une moyenne faisant entrer en ligne de compte des variations de prix pondérées par les dépenses correspondant aux quantités consommées.* Ceci ne suffit pas à le définir car il y a alors encore de multiples possibilités d'effectuer cette moyenne. En particulier, non seulement les prix ont changé entre la période de base et la période courante mais aussi *les quantités des produits consommés.* Que faut-il prendre alors comme quantités pour calculer les pondérations de l'indice de prix ? Les quantités consommées de la période de base ? Celles de la période courante ? Ou une moyenne des quantités de la période de base et de la période courante ?

La théorie des indices ne donne pas de réponse définitive et unique même dans le cas d'un seul consommateur et encore moins dans le cas de multiples consommateurs. Elle indique cependant que, sous certaines hypothèses, l'une des meilleures approximations d'un indice idéal serait l'indice de Fisher résultant d'une moyenne entre un indice basé sur les pondérations de la période de base (indice de Laspeyres) et d'un indice basé sur les pondérations de la période courante (indice de Paasche). La théorie montre aussi que, le plus souvent, l'indice de Laspeyres a tendance à surestimer l'indice de Fisher (et l'indice de Paasche à le sous-estimer). *L'idée simple derrière ce résultat est que l'indice de Laspeyres donne un poids trop important aux produits dont le prix augmente le plus, alors que ces produits vont logiquement voir leur poids diminuer dans le budget des consommateurs, dès lors que ceux-ci admettent une certaine substitution entre produits à utilité constante.*

Cependant, le calcul d'un indice de Fisher est, dans la pratique, impossible, tout au moins en cours d'année et dans les délais brefs réclamés pour un indice tel que l'IPC. D'abord, il demande la connaissance des pondérations de la période courante qui ne sont connues qu'avec des délais importants. Pour calculer l'indice de Fisher de 1998 par rapport à 1990 par exemple, il faudrait notamment pouvoir disposer des quantités consommées annuellement en 1998. Ceci n'est bien entendu pas possible *en cours* d'année 1998. Ensuite, on ne dispose tout simplement pas d'informations sur les pondérations pour les niveaux détaillés de calcul de l'indice.

C'est pourquoi la plupart des pays calculent l'IPC sous la forme d'un indice de Laspeyres, c'est-à-dire en utilisant des pondérations fixes issues de la période de base ou tout simplement, au niveau détaillé, des pondérations égales pour les produits, donc par définition fixes. Suivant les pays, l'année sur laquelle ces pondérations sont estimées est plus ou moins récente. Plus cette année sera ancienne, plus la surestimation pourrait être forte. Inversement, plus cette année sera récente plus la surestimation sera faible, voire négligeable. En France, les pondérations qui permettent d'obtenir l'indice d'ensemble à partir des indices des « postes » [4] sont mises à jour tous les ans à partir de données récentes. À l'opposé, aux Etats-Unis, la base de pondération est beaucoup plus ancienne.

Une décomposition en plusieurs niveaux

Mais la situation n'est pas aussi simple car un indice de prix aussi complexe que l'IPC résulte d'agrégations successives d'indices, chaque niveau d'agrégation ayant ses propres pondérations indépendantes, d'un « âge » variable. Ainsi en France, c'est à un niveau d'agrégation assez élevé que sont mises à jour tous les ans les pondérations à partir de données de l'année a-2. Les autres niveaux, plus détaillés, ne sont pas traités de la même façon.

Aux Etats-Unis comme en France, on peut décomposer le processus d'agrégation qui permet d'obtenir l'indice d'ensemble à partir des relevés de prix élémentaires en plusieurs étapes. La première étape, qu'on appellera « niveau détaillé », sera l'étape de calcul permettant d'obtenir les indices très détaillés (on les appelle souvent « micro-indices ») représentant des catégories de produits très fines (les « variétés ») pour une région géographique déterminée (les « agglomérations »). Ils sont obtenus à partir des relevés de prix dans les divers points de vente de l'agglomération considérée. Une deuxième étape sera le calcul de l'indice d'ensemble à partir de ces micro-indices. C'est ce que la commission Boskin a appelé « niveau agrégé ». À chacune de ces étapes, une formule de Laspeyres est utilisée (ou était utilisée en

4. Les « postes » correspondent à de grandes catégories de produit constituant le premier niveau de publication de l'IPC français. Par exemple, il y a le poste « fruits frais », ou le poste « automobiles », ou le poste «coiffeurs pour homme ». Les postes sont au nombre de 265 dans l'actuel indice, dit de « base 1990 ». Ils étaient au nombre de 295 dans l'indice dit de « base 1980 ».

France, comme on le verra). Pour le niveau agrégé, les pondérations proviennent en France de la comptabilité nationale. Dans le cas des Etats-Unis et de la plupart des autres pays européens, les pondérations reposent sur les enquêtes publiques sur les dépenses des ménages. Pour le niveau le plus détaillé, du fait de l'absence d'information, l'usage est d'accorder à chaque produit un poids égal et fixe [5]. À chacune de ces étapes, une surestimation (qu'on appellera par la suite « biais de substitution ») plus ou moins importante pourrait intervenir.

Estimations des biais de substitution

La commission Boskin a estimé qu'il y avait un biais de substitution agrégé de 0,15 % par an pour les Etats-Unis sur la base d'une structure de pondérations de l'IPC US qui remonte actuellement à 1982-1984, c'est-à-dire à plus de treize ans. Elle s'est appuyée sur des calculs de simulation effectués aux Etats-Unis qui montrent en effet que, dans les conditions actuelles, un indice de Laspeyres dont les pondérations remontent à environ dix ans augmente plus vite d'environ 0,1 % à 0,3 % par an par rapport au même indice calculé en utilisant une formule de Tornqvist (très proche de la formule de Fisher). D'où le chiffre de 0.15% l'an retenu par la commission Boskin et qui apparaît dans le tableau 1.

Quant au biais de substitution de niveau détaillé, la commission l'a estimé en comparant l'indice actuel, calculé en utilisant comme micro-indice une moyenne arithmétique des rapports de prix (correspondant à une formule de Laspeyres classique avec des pondérations égales), à un indice calculé en utilisant une moyenne géométrique. Comme nous le verrons plus bas, cette moyenne géométrique correspond, sous certaines hypothèses, dont celle d'une élasticité de substitution égale à 1, à l'IUC. A d'autres écarts près, la comparaison de ces deux indices sur plusieurs années montre une écart de 0.25% par an. D'où, le chiffre de 0.25% qui apparaît dans la deuxième ligne du tableau 1.

Mais, au delà des biais de substitution, pratiquement la moitié du biais total de 1.1% l'an avancé par la commission Boskin porte sur ce qu'elle a qualifié de biais sur les nouveaux produits. De quoi s'agit-il ?

Nouveaux produits

Une des difficultés majeures de la construction des indices de prix réside dans la contradiction qu'il y a entre la fixité des produits nécessaire au principe même de

---

5. On verra plus loin qu'avec l'informatisation de la distribution et la standardisation des codes-barres, des pondérations explicites à ce niveau de détail, qui apparaissaient de la science fiction statistique il y a quelques années, pourraient devenir réalité dans les années qui viennent.

calcul d'une comparaison des prix à deux périodes différentes et la réalité économique qui est faite d'apparition de nouveaux produits [6] et de disparition de produits obsolètes.

À chaque disparition/remplacement, le prix du produit qui remplace le produit que l'on ne retrouve plus sur les rayons doit néanmoins être comparé au prix de ce dernier. Pour évaluer la variation des prix entre l'ancien et le nouveau, il faut corriger le rapport des prix de l'éventuelle différence de qualité entre les deux produits [7]. Par exemple, si on remplace un modèle de voiture sans climatisation avec le même modèle mais équipé de la climatisation, on ne pourra bien sûr pas comparer directement leurs prix. Il faudra estimer le « prix » de la climatisation, par exemple en se fondant sur le prix qui était donné dans le catalogue du constructeur lorsque la climatisation était en option, et l'ôter du prix du nouveau modèle pour aboutir à la variation des prix « à qualité égale ». Dans cet exemple cette opération apparaît comme relativement facile, l'estimation de la valeur de l'option « climatisation» étant relativement simple. On conçoit aisément que dans d'autres cas ce traitement puisse être beaucoup plus difficile, la notion de qualité, et plus encore son estimation chiffrée, étant souvent insaisissable.

Les constructeurs d'indice de prix reconnaissent volontiers qu'il y a là une source majeure de problème. La théorie simple des indices n'est pas très éclairante puisqu'elle suppose, par définition, que les produits existent à la période de base et à la période courante. Or ce n'est précisément pas le cas. La référence de la théorie au principe de conservation du *niveau d'utilité* du consommateur permet cependant d'éclairer l'objectif que doit se fixer le statisticien. Quand un produit en remplace un autre, la variation de prix doit être calculée après avoir ramené les deux produits à un niveau d'utilité égale. Mais déjà difficile à cerner dans le cas d'un consommateur, cette notion l'est encore plus pour des millions de consommateurs. Des erreurs se produisent donc très certainement et de nombreux économistes, dont ceux de la commission Boskin, pensent que, *nos économies concurrentielles conduisant à une amélioration globale de la qualité et de la gamme des produits*, la majorité des erreurs de traitement dans l'indice des prix se produisent dans un sens, celui de la sous-estimation de l'amélioration de la qualité, et qu'il y a donc globalement une surestimation de l'inflation.

Plusieurs études approfondies, utilisant des données réelles et basées sur des méthodes rigoureuses, ont été publiées qui vont effectivement dans le sens d'un constat de sous-estimation par l'IPC de l'augmentation de la qualité et donc de surestimation de l'inflation. L'étude la plus large et la plus citée est américaine et concerne tout le secteur des *biens durables* aux Etats-Unis (Gordon, 1990). Elle

6. Pour alléger le texte, on utilisera dans ce chapitre le terme de *produits* bien qu'en toute rigueur il serait préférable de parler de « nouveaux biens et services ». Les « nouveaux services » représentent probablement la plus large part des « nouveaux produits ».
7. C'est ce que les constructeurs d'indices de prix français appellent le « traitement de l'effet-qualité ».

situait la surestimation à 1 % par an pour les années soixante-dix pour ces produits pris globalement [8]. Pour fixer les idées, les biens durables représentent 10 % de l'indice d'ensemble en France. L'impact mécanique de cette surestimation sur ce dernier serait donc de 0,1 %. En procédant à des extrapolations de ce type d'études et en se basant sur des hypothèses ad hoc qui apparaissent maintenant assez hasardeuses, la commission Boskin s'est donc lancé dans une estimation, classe de produit par classe de produit, d'un biais. Elle aboutit ainsi à un chiffre global de 0,6 % par an.

Les raisonnements faits par la commission sont le plus souvent intéressants. Cependant, nombreux sont ses critiques américains ou étrangers qui ont utilisés pour qualifier son approche le mot de « *guesstimates* ». Ainsi, certaines des estimations (sur les fruits et légumes, par exemple) n'ont pas résisté à l'analyse. D'autres estimations paraissent également exagérées. Enfin la commission a écarté les cas où des erreurs dans les méthodes actuelles pourraient conduire au contraire à sous-estimer l'inflation. Au total, le chiffre de 0,6% d'estimation du biais des nouveaux produits a concentré sur lui l'essentiel des critiques de ceux qui se sont opposés aux conclusions de la commission.

Nouveaux circuits de commercialisation

La dernière source de surestimation proviendrait d'un biais lié à un mauvais traitement des gains de parts de marché des grandes surfaces qui vendent à un prix plus bas. Aux Etats-Unis comme en France, de nouveaux circuits de distribution à prix plus bas se sont multipliés, gagnant année après année des parts de marché de plus en plus importantes aux dépens des circuits de distribution traditionnels. Il s'agit de l'essor bien connu des grandes surfaces. D'abord portant sur les super puis les hypermarchés, le mouvement a été relayé ces dernières années par l'apparition des « *hard-discounters* » et, dans le secteur des services, par les chaînes de franchisés par exemple dans le secteur de l'entretien automobile ou des travaux photographiques. Le même phénomène est apparu dans le secteur des transports aériens du fait de la dérégulation.

Or on va voir que la méthode de calcul de l'indice des prix à la consommation, aux Etats-Unis comme en France, pourrait ne pas tenir totalement compte des baisses de prix que peuvent ressentir les consommateurs d'une région ou d'un marché lorsqu'un nouveau magasin ou un nouveau producteur de service s'y installe.

Dans le cas d'un nouveau magasin, la méthode utilisée par l'IPC revient à introduire les nouveaux relevés de prix à un niveau d'indice égal à l'indice des prix des

8. Gordon aboutit à une surestimation de 1,54 % par an sur la période 1947-1983 se décomposant en 2,21 % sur la période 1947-1960, 1,24 % sur la période 1960-1973 et 1,05 % sur la période 1973-1983.

anciens relevés de cette agglomération. Par exemple, si le prix du litre de soda dans l'agglomération A était de 12F en décembre 1996 conduisant à un indice de 112,3, base 100 en 1990, et qu'une nouvelle grande surface s'était installée dans cette même agglomération dans laquelle on relevait pour la première fois un prix de 8F le litre à la même période, le niveau de départ de l'indice élémentaire correspondant au soda dans cette nouvelle grande surface sera aussi égal 112,3. Sa fusion avec les autres indices de l'agglomération ne conduira donc pas à une baisse de l'indice de prix du soda dans l'agglomération [9]. L'indice n'enregistrera une variation que si les petits commerçants (ou les autres moyennes ou grandes surfaces) dont on suivait les prix auparavant baissaient eux-mêmes leur prix du fait de la concurrence de la nouvelle grande surface.

Tout se passe donc en fait comme si les statisticiens considéraient que, à produit égal, la totalité de la différence de prix entre les deux circuits de distribution était en quelque sorte due à une différence de qualité du service commercial. Il est vrai que les actes d'achat dans un commerce traditionnel et dans une grande surface ne sont pas équivalents même dans le cas où le produit vendu serait strictement le même. La proximité du lieu de résidence, les services personnalisés rendus au client, la convivialité ont été souvent cités en faveur des circuits traditionnels. L'essor des grandes surfaces ne s'explique d'ailleurs pas seulement par des prix plus bas. Il est largement lié à la civilisation de l'automobile, au développement des banlieues et à l'équipement des ménages en congélateurs, tous phénomènes permettant des achats groupés et importants et correspondant à un service commercial différent. Cependant, l'acuité de la concurrence et des « guerres de prix » entre circuits de commercialisation, que traduisent les gains continus de parts de marché des grandes surfaces, permettent tout aussi sûrement de penser que l'hypothèse implicite des statisticiens revenant à considérer que la totalité de la différence de prix s'explique par la différence de service est exagérée.

Il y a donc là une source évidente de surestimation de la hausse des prix tout simplement par omission des baisses de prix liées au développement des grandes surfaces. Le traitement statistique approprié consisterait à pouvoir estimer la valeur que le consommateur accorde à un déplacement de ses achats d'un type de commerce à un autre. Certaines études ont été faites aux Etats-Unis sur ce sujet mais n'ont pas abouti encore à des procédures opérationnelles. Une proposition pourrait être de considérer que la moitié de la différence de prix entre les circuits de

---

9. Il est à noter que le fait que l'indice des prix du soda ne baisse pas va entraîner une différence importante entre l'évolution du « volume » de vente *(au sens de la comptabilité nationale)* de soda dans cette agglomération au moment de l'installation de la grande surface et l'évolution du nombre de litres vendus. Le traitement de l'apparition de la nouvelle grande surface dans l'indice des prix implique en effet que l'on considère qu'un litre de soda en grande surface est moins « bon (?) » pour le consommateur et donc « pèse » moins dans le volume total vendu qu'un litre vendu dans un circuit traditionnel, plus cher. L'idée est que le service commercial associé à l'achat de soda en grande surface est moindre que dans le circuit traditionnel.

commercialisation est une différence de prix et l'autre moitié une différence de service. Mais ceci peut paraître aussi arbitraire que l'hypothèse actuelle.

L'étude la plus approfondie sur la question de l 'impact sur l'indice des prix des gains de parts de marché des grandes surfaces est une étude française publiée en 1995 (Saglio, 1995 ; Prime et Saglio, 1995 ; Dubeaux et Saglio, 1995). Dans cette étude, qui est une extrapolation d'une monographie très détaillée sur le cas des tablettes de chocolat, la différence entre un indice calculé suivant la méthode traditionnelle et un indice qui considérerait, à l'inverse, que la totalité de la différence de prix entre circuits de distribution est une différence « pure » de prix, est estimée à 0,2 % l'an pendant les années quatre-vingt. Ce chiffre de 0,2 %, appelé « effet circuit d'achat », constitue donc probablement un majorant du biais dû aux nouveaux circuits de distribution, si l'on admet qu'une partie au moins de la différence de prix s'explique par une différence de service commercial. Si l'on admet notamment l'hypothèse que seule la moitié de la différence est une différence de prix, le biais se limiterait ainsi à 0,1 % par an.

Peut-on considérer ce chiffre de 0,1 % dont la base de départ a été estimée sur les années quatre-vingt comme représentatif pour les années futures ? D'un côté, il est probable que le développement très rapide des grandes surfaces soit maintenant ralenti. En France, la récente loi qui rend plus difficile la création de nouvelles grandes surfaces va dans ce sens. Ceci tendrait donc à retenir un chiffre inférieur. D'un autre côté, l'estimation originale de 0,2 % pourrait être elle-même sous-estimée car elle n'a pas totalement pris en compte le phénomène des nouveaux circuits de distribution et de la dérégulation dans le secteur des services, des transports et des télécommunications. Le développement des chaînes de franchisés dans les secteurs de l'entretien automobile et des travaux photographiques en constituent des exemples frappants de même que l'effet spectaculaire sur les prix et les parts de marché de la dérégulation dans le transport aérien domestique. C'est ainsi que, dans le secteur du transport aérien en 1993, la comptabilité nationale française a estimé un indice de prix sensiblement plus faible que l'indice des prix à la consommation correspondant. Il y a encore aussi de larges réserves de « guerres de prix » par de nouveaux entrants dans les secteurs de la banque et de l'assurance par exemple avec la banque et l'assurance par téléphone et sans parler des possibilités d'achat à distance par Internet. Au total, on ne peut donc pas écarter la possibilité d'une légère surestimation, due à une prise en compte incomplète des nouveaux circuits de commercialisation, comprise dans une fourchette de 0,05 à 0,15 % par an pour la France. Ces chiffres recoupent celui retenu par la commission Boskin pour les Etats-Unis.

2. Un débat très chaud....qui s'est calmé apparemment très vite...

Malgré toutes les critiques qui ont suivi, il faut d'abord reconnaître que la commission Boskin a réussi à garder à son rapport un bon niveau scientifique. L'annonce de ses conclusions n'en a pas moins fait l'effet d'une bombe à l'aune des reprises traditionnelles des médias sur la statistique. Pour une fois les médias américains ont braqué leurs projecteurs sur la statistique, dans un climat qui est devenu rapidement très polémique,. Il faut dire que l'enjeu politique était de taille puisque le Sénat menaçait de corriger l'IPC de son biais pour faire diminuer les indexations des diverses dépenses fédérales et ainsi parvenir à équilibrer plus facilement le budget. Le président de la FED, Alan Greenspan, un des inspirateurs de la commission Boskin, a d'ailleurs bataillé ferme au Congrès pour que ses conclusions soient retenues. On en était à deux doigts.

Et puis, deux phénomènes se sont produits. D'une part, le BLS, soutenu par une communauté grandissante de statisticiens américains et étrangers, a fait valoir que les conclusions de la commission n'étaient pas si incontestables que cela et qu'il ne fallait pas se précipiter. Ensuite, les hommes politiques américains ont visiblement découvert à ce moment qu'ils avaient finalement d'autres priorités que celles d'intervenir dans la mesure de la variation des prix. La raison principale en est probablement que tout le monde s'est aperçu alors qu'il n'y avait pas besoin de corriger l'IPC pour parvenir à un équilibre du budget fédéral. La croissance américaine se porte si bien que les rentrées fiscales font exploser les prévisions les plus optimistes et que les nouvelles simulations à moyen terme du déficit montrent qu'il sera équilibré sans aucune mesure spécifique. Se rajoutant à cela, certains hommes politiques se sont rendu compte que d'intervenir sur la mesure du coût de la vie pouvait ne pas être très rentable électoralement. Cela fait perdre les voix des pauvres dont les prestations sont indexées sur l'IPC. Cela peut aussi faire perdre les voix des classes moyennes, dans la mesure où les tranches du barème de l'impôt sur le revenu sont également indexées sur l'indice des prix. Moins d'indice des prix signifie que plus de gens passent dans le barème supérieur, ce qui n'est évidemment pas populaire, surtout aux Etats-Unis. Au total, la classe politique s'est apparemment encore plus rapidement désintéressée du sujet qu'elle ne l'avait saisi. Mais ce n'est pas le cas d'Alan Greenspan qui a fait une intervention remarquée sur les problèmes de mesure de l'inflation à un récent congrès d'économie aux Etats-Unis (Greenspan, 1998). Le rapport Boskin n'a donc pas eu de conclusion institutionnelle. Cela ne doit en aucune façon diminuer son mérite principal qui est d'avoir relancé les études sur la mesure de la variation des prix. En passant, il a permis d'ailleurs au BLS d'obtenir des crédits de recherche plus importants !

Les instituts de statistique étaient évidemment les plus concernés par les conclusions de la commission. Leur réactions « moyennes » peut être résumée ainsi. Une assez grande unanimité s'est formée pour reconnaître l'existence d'un danger de biais de

substitution agrégé. Cependant de nombreux pays, dont la France, ont souligné que la situation américaine (pondérations vieilles de plus de dix ans) était exceptionnelle. Une des solutions consiste évidemment à mettre à jour beaucoup plus rapidement les pondérations, ce que beaucoup de pays font. Une assez grande unanimité s'est aussi formée pour reconnaître le danger du biais de substitution détaillé. Par contre, les solutions sont encore discutées. Les pays européens ont introduit récemment la moyenne géométrique dans leur indice au niveau le plus détaillé. Le BLS hésite toujours à le faire et propose des pistes de recherche plus sophistiquées. On y reviendra plus bas. La réalité d'un biais de circuits de commercialisation est également relativement partagée par les statisticiens, mais son ampleur est discutée et peu de solutions concrètes y sont apportées.

Par contre, la plupart des instituts de statistique qui se sont exprimés se sont montrés très sceptiques sur l'évaluation de 0.6% du biais de nouveaux produits. Ce chiffre leur paraît très exagéré. Cependant, tous reconnaissent que les méthodes pratiques d'évaluation du changement de qualité des produits sont largement en dessous de ce qui est souhaitable et tous fixent comme priorité d'action des études dans ce sens. Nous en verrons des exemples plus bas. Enfin, l'une des pistes des plus fécondes d'études, d'ailleurs recommandée par la commission Boskin, apparaît être l'utilisation des données en provenance de panels de distributeurs. Nous en verrons également quelques exemples.

La réaction de l'Insee

L'Insee a publié un article complet qui a été largement commenté en France (Lequiller, 1997). Ses messages essentiels étaient les suivants. En premier lieu, l'IPC français n'est pratiquement pas sujet au biais de substitution agrégé, car il utilise des pondérations de l'année n-2, très proche des pondérations courantes. C'est ce qui explique la valeur nulle de la première ligne du tableau 2. Ensuite, l'utilisation de la moyenne géométrique et l'abandon de la moyenne arithmétique des rapports de prix depuis le début 1997 pour un grand nombre de variétés permet également de le mettre à l'abri du biais de substitution détaillé. C'est ce qui explique la valeur nulle de la troisième ligne du même tableau 2. Par contre, un biais résiduel faible (entre 0.05 et 0.10%) pourrait subsister à un niveau défini comme « intermédiaire ». Par ailleurs, l'article admettait également un biais de circuit de commercialisation compris entre 0.05 et 0.15%. On en a vu l'explication plus haut. Enfin, il concluait qu'il n'était pas possible de donner une estimation d'un biais quelconque pour les nouveaux produits. Il ajoutait qu'il lui semblait, qu'en tout état de cause, le 0.6% du rapport Boskin était largement surestimé.

Tableau 2 : France : surestimation de l'indice des prix à la consommation

Sources de surestimation	France
	Valeur estimée en % annuel pour les années postérieures à 1996
Substitution au niveau agrégé	-
Substitution au niveau intermédiaire	0,05-0,10
Substitution au niveau détaillé	-
Nouveaux circuits de distribution	0,05-0,15
Total hors nouveaux produits	**0,10-0,25**
Nouveaux produits	?
Total y compris nouveaux produits	?

La réaction du BLS

La réaction du BLS américain a fait l'objet d'une note officielle publiée en juin 1997 (BLS, 1997). En premier lieu, le BLS confirme qu'il accepte le cadre théorique de l'IUC, indice à utilité constante (voir Lequiller, 1997) ainsi que la commission Boskin le recommandait. Cette prise de position « conceptuelle » n'est pas nouvelle aux Etats-Unis, mais elle était moins affirmée auparavant. Elle est souvent encore contestée en Europe. Mais le BLS fait remarquer qu'entre le cadre théorique et l'application pratique de celui-ci, il y a une grande marge. Le calcul d'un indice de Fisher, qui est l'objectif affiché, est ainsi impossible sauf avec un retard de plusieurs mois sinon années. Le BLS refuse toujours le principe d'un indice de Laspeyres chaîné, apparemment car il n'a pas de fondement théorique absolu. Il est vrai qu'il n'y a pas, en toute rigueur, de relation simple entre un indice de Laspeyres chaîné et l'indice de Fisher (Greenlees, 1997). Cependant, le BLS reconnaît implicitement qu'il doit mettre à jour ses pondérations plus rapidement qu'auparavant. Il admet par ailleurs le montant du biais de substitution agrégé de 0.15%. Il récuse par contre l'estimation de 0.25% de biais de substitution détaillé. L'argument, sur lequel nous reviendrons, est que la moyenne géométrique n'est pas forcément la formule de référence puisqu'elle présuppose une élasticité de substitution de 1, qui est une hypothèse parmi d'autres, forte pour certains produits, faibles pour d'autres. Nous verrons plus bas les propositions intéressantes que le BLS fait en faveur de formules plus souples. Le BLS conteste également le montant de 0.1% du biais de nouveaux circuits de commercialisation. Il affirme que les données sur lesquelles la commission s'est fondé surestimaient le problème.

Mais la réaction la plus forte concerne le biais de nouveaux produits. Dans une analyse détaillée des estimations faites par la commission, le BLS met en avant de

nombreuses imperfections et confirme qu'il y a eu « un biais sur le biais ». D'abord, sur 19 des catégories de produits pour lesquelles la commission avait estimé qu'il y avait un biais, il fait remarquer que l'estimation repose sur des a priori subjectifs. Pour deux exemples précis (fruits et légumes, essence) pour lesquels le BLS a trouvé des données permettant d'aller au-delà de cette subjectivité, il trouve des résultats bien inférieurs à ceux de la commission. Pour quatre autres catégories importantes (les loyers, les vêtements, les voitures neuves et d'occasion) les données utilisées par la commission apparaissent viciées (Moulton-Moses, 1997, Triplett, 1997). Par contre, pour le reste des produits (essentiellement les biens durables « high-tech » et les biens et services liés à la santé), le BLS reconnaît que la commission s'est appuyé sur des sources sérieuses et que ses critiques sont plus fondées.

Cette critique de la critique de la commission s'accompagne d'un vigoureux programme de recherche dont les axes sont multiples: publication d'indices expérimentaux utilisant la moyenne géométrique (1999) et une formule d'indice « superlatif » (publication officielle en 2002); mise à jour plus fréquente des pondérations; tests de nouvelles formules de micro-indices; introduction de méthodes nouvelles de prise en compte de l'augmentation de la qualité (régression hédoniques, prix de services « groupés » pour les services hospitaliers et médicaux).

Les autres instituts de statistique

Les instituts anglais et canadien se sont également exprimés sur la question (Ducharme, 1997). Les Canadiens soulignent que le biais de leur IPC devrait être beaucoup plus faible que celui des Etats-Unis par le fait que , comme en France, certaines des méthodes de calcul introduites dans les années récentes avaient devancé les critiques faites par la commission Boskin. Ainsi au Canada les pondérations sont mises à jour tous les 4 ans, au lieu de tous les dix ans aux Etats-Unis. Stat Can annonce par ailleurs qu'ils vont passer à une mise à jour annuelle. La moyenne géométrique a été généralisée depuis 1995. Les nouveaux produits sont introduits beaucoup plus rapidement qu'aux Etats-Unis. Comme les statisticiens français et américains, les Canadiens contestent l'ampleur du biais de nouveaux produits et affirment qu'il n'est pas possible de donner un chiffre. Ils annoncent des initiatives pour l'introduction de méthodes hédoniques notamment pour les produits de l'habillement.

Tableau 3 : Canada : surestimation de l'indice des prix à la consommation

Sources de surestimation	Canada
	Valeur estimée en % annuel pour les années postérieures à 1996
Substitution au niveau agrégé	0,10-0,20
Substitution au niveau intermédiaire	0,00-0,10
Substitution au niveau détaillé	-
Nouveaux circuits de distribution	0,0-0,10
Total hors nouveaux produits	**0,10-0,40**
Nouveaux produits	?
Total y compris nouveaux produits	?

Pour l'ONS britannique (Fenwick, 1997), la réaction au rapport Boskin est beaucoup plus prudente encore. D'emblée, le cadre de référence à l'indice à utilité constante est rejeté. Comme en France, les pondérations de l'indice de prix britannique sont mises à jour tous les ans réduisant sérieusement le risque de biais de substitution agrégé. Concernant les autres biais, l'ONS, toujours aussi très prudent demande des études préalables sur la moyenne géométrique et l'introduction de nouvelles méthodes d'estimation de la qualité avant de prendre une quelconque décision.

3. Quatre pistes d'études méthodologiques

La plupart des critiques ont porté sur l'ampleur du biais tel qu'il était estimé par la commission. Mais ceci ne doit pas masquer que, en fait, nombreux sont les statisticiens qui admettent l'existence des problèmes de fond qui ont été signalés par la commission et qui préconisent de relancer très sérieusement des initiatives pour améliorer les méthodes des indices de prix. Quatre pistes essentielles semblent prendre corps. La première porte sur les formules d'indice. La seconde sur les régressions dites « hédoniques ». La troisième sur la notion de service groupé. La quatrième sur l'utilisation des données dites « scanner ».

Formules d'indice

Les principaux résultats à rappeler sont les suivants: le cadre théorique de référence est la théorie de l'indice à utilité constante (IUC); on s'efforce d'approximer l'IUC;

l'indice de Fisher en est une bonne approximation. Dans ce cadre, on peut prouver aussi que dans le cas où les courbes d'utilité du consommateur ont une forme Cobb-Douglas ($Aq_1^a q_2^b$), la moyenne géométrique pondérée par les valeurs des dépenses est la forme exacte de l'IUC. Transplanté dans le cas des micro-indices, dans lequel les dépenses sont implicitement égales pour tous les produits, on en déduit que la moyenne géométrique simple est égale à l'IUC si les fonctions d'utilité sont des Cobb-Douglas.

On peut retrouver ces deux derniers résultats facilement.

Soit le programme de maximisation sous contrainte: $Max\ u(q_1, q_2) = Aq_1^a q_2^b$, sous la contrainte $y = p_1 q_1 + p_2 q_2$. La différentiation du Lagrangien conduit aux équations suivantes :

$$a = \frac{p_1 q_1}{y}\ et\ b = \frac{p_2 q_2}{y}.$$ [10]

On en tire :

$$y = \left(\frac{p_1}{a}\right)^a \left(\frac{p_2}{b}\right)^b u(q_1, q_2).$$

L'IUC étant égal au ratio du budget correspondant au nouveau vecteur de prix *à utilité constante* sur le budget initial, on peut donc écrire :

$$I_U = \frac{y(1)\ \text{à}\ utilit\acute{e}\ \text{constante}}{y(0)},\ \text{soit}$$

$$I_U = \frac{\left(\frac{p_{1.1}}{a}\right)^a \left(\frac{p_{2.1}}{b}\right)^b u(q_1, q_2)}{\left(\frac{p_{1.0}}{a}\right)^a \left(\frac{p_{2.0}}{b}\right)^b u(q_1, q_2)}$$

, qui, parce que les valeurs de la fonction *u* au numérateur et au dénominateur sont égales par construction, se simplifie en :

[10] On gardera en mémoire ce résultat intermédiaire qui indique que, par construction, pour une fonction d'utilité de type Cobb-Douglas, la pondération de chaque produit dans les dépenses totales du consommateur est égale à a (ou b) et *reste donc fixe*.

$$I_U = \left(\frac{p_{1.1}}{p_{1.0}}\right)^a \cdot \left(\frac{p_{2.1}}{p_{2.0}}\right)^b$$, qui est la moyenne géométrique, CQFD.

La moyenne géométrique pondérée par la part de chaque produit dans les dépenses du consommateur est donc l'indice IUC pour une fonction d'utilité de type Cobb-Douglas. On se rappelle que pour une Cobb-Douglas la part d'un produit dans la dépense ne change pas (en d'autres mots, la pondération en valeur ne change pas). Tout se passe donc comme si, lorsque les consommateurs perçoivent un changement dans les prix relatifs, ils substituent les quantités de produits de façon à garder les parts dans les dépenses fixes.. Le produit qui augmente le plus verra sa consommation diminuer, le produit qui augmente le moins sa consommation augmenter. L'élasticité de substitution est égale à 1. Cette hypothèse n'est certainement pas vérifiée dans le long terme pour les grands postes de la consommation puisque l'on sait bien que, par exemple, le poids des services s'est accru significativement dans l'indice alors que leur prix s'est accru plus vite que la moyenne. Par contre, au niveau des micro-indices et dans le court terme, pareille hypothèse apparaît plus réaliste que la totale fixité des quantités, typique de l'indice de Laspeyres[11].

Mais si ce raisonnement permet de justifier l'utilisation de la moyenne géométrique, il ne l'autorise que sous l'hypothèse d'une élasticité de substitution particulière, ici égale à 1. Si l'élasticité de substitution est nulle (cas de biens complémentaires), c'est plutôt l'indice de Laspeyres qu'il faudrait utiliser. Comment choisir entre l'un et l'autres ? L'idée la plus rigoureuse consisterait donc à trouver une formule de calcul avec un paramètre prenant en compte le degré de substituabilité. Une proposition récente a été faite par Moulton (Moulton, 1996).

S'appuyant sur une classe assez large de fonctions de coût de type CES (à élasticité de substitution constante), $c(u,p) = u\left[\sum a_k^{\sigma} p_k^{1-\sigma}\right]^{1/(1-\sigma)}$, il montre, suivant un raisonnement analogue à celui ci-dessus, que l'IUC correspondant à cette fonction d'utilité s'écrit:

$$I_{IUC} = \left[\sum_k s_{0k} (p_{tk} / p_{0k})^{1-\sigma}\right]^{1/(1-\sigma)},$$

[11] On peut prouver également assez facilement, et même graphiquement, que l'indice de Laspeyres est égal à l'IUC lorsque les courbes d'indifférence sont de type « Léontieff », i.e., sans substitution aucune.

où s_{0k} est la part de la dépense en produit k à la période de base. L'idée serait donc d'utiliser cette formule. Il ne resterait plus alors qu'à estimer pour chaque classe de produits l'élasticité de substitution σ. Les données pour l'estimer existent pour une partie au moins de la consommation des ménages, sous la forme des données scanner. Mais savoir si on peut résumer les substitutions à l'intérieur d'une classe de produits représentant des dizaines de produits différents en un chiffre unique est une question qui reste à explorer.

Les régressions hédoniques

Sous ce vocable savant, il s'agit simplement des méthodes de traitement des changements de qualité reposant sur l'économétrie. Dans la plupart des cas aujourd'hui, l'hypothèse qui est faite pour remplacer un produit par un autre est que la différence de prix au moment où les deux produits sont observés sur le marché représente la différence de qualité. Cette hypothèse est particulièrement simpliste. Dans l'exemple des ordinateurs où, souvent, les améliorations de la qualité sont accompagnés d'une baisse des prix, plusieurs études ont montré que cela conduisait très certainement à une sous-estimation de la baisse des prix. Les méthodes hédoniques consistent à décomposer le prix d'un produit dans les prix de ces principales caractéristiques en s'appuyant sur un modèle économétrique. En effet, dès qu'il y a plus d'une caractéristique, l'économétrie est indispensable. La méthode n'est pas nouvelle, elle date de plusieurs décennies. En France, elle est utilisée couramment pour estimer l'indice des prix des micro-ordinateurs à la production (PVI) depuis le début des années 90. Elle avait d'ailleurs été utilisée dans l'IPC français dans les années 60 et 70 mais avait été abandonnée depuis, plus faute de moyens apparemment que du fait d'une décision méthodologique.

Le rapport Boskin a relancé les études dans ce domaine. Ainsi en France, plusieurs études ont été conduites récemment dans la division Prix à la consommation. L'une porte sur les lave vaisselles et l'autre, encore à un stade exploratoire, sur l'habillement. D'après les notes internes de l'Insee (Bascher, 1997), le modèle sur les lave-vaisselle se révèle concluant. Le prix du lave-vaisselle est estimé en fonction du nombre de programmes, du nombre de températures, du degré de bruit, de la catégorie du producteur, du type de point de vente. Il apparaît sur les quelques mois d'utilisation de cette nouvelle méthode qu'elle donne des résultats peu différents de la méthode utilisée auparavant. La période de test est cependant trop courte pour porter un jugement définitif.

Le domaine de l'habillement est probablement le plus intéressant car les méthodes actuelles y sont particulièrement critiquables (Lequiller, 1997). La différence de prix entre deux produits observés pratiquement à un an d'intervalle est souvent complètement annulée, étant affectée de fait implicitement à un discutable effet de mode. Cependant, le domaine est évidemment très complexe. Il y a cependant au

moins des espoirs de voir des progrès s'accomplir puisque les études sont menées d'emblée sur un plan international avec un groupe de travail européen incluant la France, la Suède, la Finlande et le Royaume-Uni. Une des idées à la base de ces groupes de travail était qu'on pouvait envisager un partage des tâches entre pays, compte tenu de la mondialisation de plus en plus grande des marchés. La recherche progresse en France puisqu'un modèle hédonique pour les chemises pour hommes a été estimé (Bascher, 1997). Ses résultats ont été comparés avec ceux des autres pays et on envisage sa mise en application.

Outre les problèmes ardus d'estimation statistique et le coût de constitution des la base de données, une des difficultés, souvent sous-estimée, réside dans la mise en pratique des résultats de ces modèles. En France, pour les biens durables, tous les remplacements de produits sont centralisés par l'équipe parisienne chargée de l'estimation des effets-qualité qui peut donc intervenir pratiquement en temps réel sur les choix statistiques faits par les enquêteurs. Pour l'habillement, un tel système n'existe pas et la tendance actuelle serait plutôt l'inverse. Une voie à explorer réside peut-être dans les matrices de remplacement ainsi que proposé par un statisticien anglais dans le cas des téléviseurs (Silver, 1997). Ces matrices, estimées centralement, seraient utilisées par les enquêteurs sur le terrain pour calculer les prix des produits de remplacement.

Le groupage des services médicaux

Depuis un certain temps, certains économistes remettent en question les modalités même de suivi de certains produits-services. L'exemple le plus parlant est celui des services médicaux. La question est simple. Faut-il suivre séparément, avec un poids fixe pour chacune de ces prestations, le prix de la chambre d'hôpital et de l'appendicectomie ou faut-il suivre le prix de l'ensemble de la prestation ? Les résultats sont évidemment très différents. Au cours des dernières années, les techniques opératoires sont devenues de plus en plus légères permettant une convalescence à l'hôpital plus réduite. Ceci a contribué à faire baisser très sensiblement le prix global de la prestation d'ensemble. Il est clair qu'un indice de prix construit à partir d'une prestation décomposée ne reflétera pas la même baisse de prix. Il se peut même que chacun des morceaux de la prestations globale ait connu une stabilité sinon une hausse de son prix. Auquel cas, malgré toute mise à jour régulière des pondérations, l'indice montera ou stagnera mais ne baissera pas.

Les constructeurs de l'IPC américain ont opté, depuis janvier 1997, pour la voie de la prestation globale pour les services hospitaliers. Le prix suivi est celui d'une facture globale d'un patient subissant une intervention d'un certain type. Ce suivi va permettre au BLS de repérer tout changement dans le volume du traitement au cours du temps et de le traiter en conséquence. Pour le moment les décisions des autres instituts de statistique n'ont pas été rendues publiques.

Les données « scanner »

Depuis longtemps, des sociétés d'études de marché réunissent des données très détaillées sur les prix de vente et les quantités consommées de produits de grande consommation. Ces données leur servent pour les études (très chèrement payées) que les industriels ou les distributeurs commandent pour mettre au point leurs produits, suivre l'effet des campagnes publicitaires ou d'autres formes de promotion. Aujourd'hui, avec les possibilités de l'informatique, certaines de ces sociétés réunissent de manière routinière toutes les données de caisse (c'est à dire prix moyens et quantités vendues) de plusieurs centaines d'hypermarchés et supermarchés pour tous les produits vendus sous forme de code-barre. Ces données représentent un fond exceptionnel, proche de l'univers (bien qu'il ne couvre pas les petits circuits de commercialisation), pour un ensemble significatif de la consommation des ménages (estimé à 13% aux Etats-Unis).

La société Nielsen a proposé aux instituts de statistique de plusieurs pays (France, Canada, Etats-Unis, Suède, Pays-Bas) des extraits gratuits de ces données détaillées à fins d'études. Ces dernières sont très prometteuses, d'autant plus, qu'avec les progrès en rapidité de mobilisation des données, on ne peut plus rejeter l'hypothèse que, dans un futur pas si lointain, cette source de données pourrait être utilisée de manière courante pour le calcul des IPC. Ces données ne seraient évidemment pas gratuites. Mais pour le moment aucune étude sérieuse de prix n'a encore été faite.

La première observation exprimée par tous ceux qui ont manipulé ces données est la masse énorme que cela représente. Même les disques durs des plus gros micros que l'on possède ont du mal à stocker les données pour quelques produits pour quelques mois! L'univers, dans sa complexité, éclate au grand jour ! Une fois cette difficulté levée (la règle du 20-80, 20% des produits représente 80% des ventes, est bien utile), les résultats sont riches.

La première piste d'étude porte très naturellement sur l'apport révolutionnaire que représente pour les constructeurs d'indices de prix les *quantités consommées en relation avec les prix observés*. A partir de ce moment, on peut calculer toutes sortes d'indices mensuels (ou même hebdomadaires) « vrais » au niveau détaillé, en lieu et place des micro-indices très pauvres qui sont actuellement utilisés: Laspeyres, Paasche, Fisher, sous formes directes ou chaînées, valeurs unitaires (sur les produits, sur les magasins). Les études concluent nettement en défaveur du Laspeyres chaîné (tout à fait catastrophique au niveau mensuel détaillé). Le Fisher se situe toujours nettement en dessous du Laspeyres, ainsi que la théorie le prévoyait. Le Fisher chaîné apparaîtrait comme une bonne option. (Hawkes, Dalen, Bradley-Cook-Leaver-Moulton, de Haan-Opperdoes, 1997).

Le calcul de valeurs unitaires au lieu d'indices de prix permet d'analyser des « effets circuits d'achat » qui se confirment comme importants. On peut faire des moyennes

pour un produit très précis donné sur différents magasins (effet circuit d'achat), ou des moyennes par magasin pour des produits peu différents (traitement des nouveaux produits en variété homogène). Mais, bien que permettant toutes les analyses sur l'impact sur les indices de prix des gains de part de marchés dues à des promotions ou à des guerres de prix entre grandes chaînes de distribution, ces données ne permettent cependant pas d'analyser les gains sur d'autres circuits de commercialisation (marchés, magasins traditionnels) qui ne font pas partie de l'échantillon. Ces données ne permettent pas non plus de suivre systématiquement les effets de l'ouverture d'un nouveau magasin dans une région.

Par ailleurs, a méthodes constantes, ces données permettent de mettre à jour beaucoup plus rapidement les pondérations annuelles de l'indice des prix, y compris, en y ajoutant des données de panels de consommateurs, sur les parts de marchés des circuits de commercialisation. Le biais de substitution « intermédiaire » pourrait ainsi être résolu. Ceci est en bonne voie en France.

La deuxième grande voie d'étude porte sur la réduction des erreurs d'échantillonnage par la mise en place de procédures plus sophistiquées de choix d'échantillon et tout simplement par accroissement de la taille de l'échantillon (cependant de Haan-Opperdoes font remarquer que l'échantillon « scanner » est beaucoup plus riche en terme de produits mais pas en terme de magasins). Des procédures de sélection probabiliste de variétés en fonction de la taille des ventes peuvent être mises en place à faible coût. (Bradley-Cook-Leaver-Moulton, de Haan-Opperdoes, 1997). Scobie (1996) fait apparaître que la richesse des données permet d'intégrer des produits non standards (marques peu connues, volumes unitaires plus importants) que les IPC ont tendance à exclure par simplification.

Enfin, la dernière voie d'étude porte sur l'utilisation de ces données pour repérer les nouveaux produits, accélérer leur introduction dans l'indice, et améliorer cette intégration (traitement de l'effet qualité). Le repérage des nouveaux produits et de leur poids est rendu évidemment beaucoup plus aisé. Chose moins connue, ces bases de données contiennent des caractéristiques précises pour les produits permettant d'envisager l'utilisation de méthodes hédoniques sur une beaucoup plus grande échelle qu'aujourd'hui (la barrière à l'entrée pour les méthodes hédoniques est le coût de constitution de la base de données), (Silver, 1997).

BIBLIOGRAPHIE

BAKER D., « The Overstated CPI Can It Really Be True », *Challenge,* sep./oct., pp. 26-33, États-Unis, 1996

BASCHER J. « Le modèle hédonique français sur la chemise: premiers résultats et comparaisons internationales », *mimeo Insee* (note 429/F320 du 26/12/97), 1997

BRADLEY R., COOK B., LEAVER S., MOULTON B., « An Overview of Research on Potential Uses of Scanner Data in the US CPI », *Third meeting of the International Group on Price Indices*, Voorburg, Netherlands, April 16-18, 1997

DE HAAN J., OPPERDOES E., « Estimation of the Coffee Price Index using Scanner Data », *Third meeting of the International Group on Price Indices,* Voorburg, Netherlands, April 16-18, 1997

DUCHARME, L-M., « L'IPC Canadien et la question des biais: le présent et l'avenir », dans « *Biais de l'IPC: les expériences de cinq pays de l'OCDE* », Série analytique de la Division des prix, Statistique Canada, 1997

BLS, « Measurement Issues in the Consumer Price Index », *Site Internet du BLS (stats.bls.gov/cpihome.htm),* Juin 1997

BOSKIN M., DULBERGER E., GRILICHES Z., GORDON R. ET JORGENSEN D. « *Toward a More Accurate Measure of the Cost of Living* », Final Report to the Senate Finance Committee, décembre, États-Unis, 1996

DALEN J., « Experiments with Swedish Scanner Data », *Third meeting of the International Group on Price Indices*, Voorburg, Netherlands, April 16-18, 1997

DUBEAUX D. ET SAGLIO A.,« Modification des circuits de distribution et évolution des prix alimentaires », *Économie et Statistique,* n° 285-286, pp. 49-58, 1995.

FENWICK D., « The Boskin Report from a United Kingdom Perspective », dans « *Biais de l'IPC: les expériences de cinq pays de l'OCDE* », Série analytique de la Division des prix, Statistique Canada, 1997

GORDON R., *The Measurement of Durable Goods Prices*, University of Chicago Press for the NBER, États-Unis, 1990

GREENSPAN A., « Remarks by Chairman of the Board of the US FED », Annual Meeting of the American Economic Association and American Finance Asociation, Chicago, 3/1/1998.

HAWKES W., « Reconciliation of Consumer Price Index Trends with Corresponding Trends in Average Prices for Quasi-Homogeneous Goods using Scanning Data », *Third meeting of the International Group on Price Indices*, Voorburg, Netherlands, April 16-18, 1997

GREENLEES J., « Expenditure Weight Updates and Measure Inflation », *Third meeting of the International Group on Price Indices*, Voorburg, Netherlands, April 16-18, 1997

LEQUILLER, F., « L'indice des prix surestime-t-il l'inflation ? », *Economie et Statistique*, n°303, Insee, 1997.

MOULTON B, « Constant Elasticity Cost-of-Living Index in Share-Relative Form », *BLS mimeo*, 1996

MOULTON B., MOSES K., « Addressing the Quality Change Issue in the Consumer Price Index », *Brookings Papers on Economic Activity*, 1997

PRIME M ET SAGLIO A., « Indices de prix et prix moyens : une étude de cas », *Économie et Statistique,* n° 285-286, pp. 35-48, 1995.

SAGLIO A., « Changement de tissu commercial et mesure de l'évolution des prix », *Économie et Statistique,* n° 285-286, pp. 9-33, 1995.

SCOBIE H., « Potential Uses of Scanner Data in the Production of Price Indexes- A case Study using Coffee Data », *mimeo Statistique Canada*, 1996

SILVER M, IOANNIDIS C., HAWORTH M., « Hedonic Quality Adjustments for Non Comparable Items for CPIs », *Third meeting of the International Group on Price Indices*, Voorburg, Netherlands, April 16-18, 1997

TRIPLETT J., « The Post 73 Consumption Slump: Myth or Reality ? », *The Federal Reserve of Saint Louis Review*, 1997

Session 5

Faire pour apprendre : trois expériences de formation active aux enquêtes

LA RÉALISATION D'UNE ENQUÊTE À L'ÉCOLE NATIONALE DE LA STATISTIQUE ET DE L'ANALYSE DE L'INFORMATION

Michel Simioni, Yves Tillé

La méthodologie d'enquête à l'Ensai

L'Ecole Nationale de la Statistique et de l'Analyse de l'Information (Ensai) s'est installée sur le Campus de Ker Lann, commune de Bruz, à 8 kilomètres de Rennes en septembre 1996. Après un recrutement sur concours au niveau bac+2, les élèves de l'Ensai suivent une scolarité de trois ans s'ils ne sont pas fonctionnaires (élèves dits «titulaires») et une scolarité de deux ans s'ils sont fonctionnaires (attachés) à l'Institut de la Statistique et des Etudes Economiques (Insee). Les élèves fonctionnaires complètent leur scolarité par la formation continue diplômante des attachés. Cette formation consiste en 20 modules d'enseignement d'une durée d'une semaine. Pour les élèves titulaires, la troisième année est une filière de spécialisation. L'Ensai offre actuellement un choix de quatre filières : statistique pour les sciences de la vie, statistique pour l'industrie, systèmes d'information statistique, et statistique pour l'économie, les sciences sociales et la gestion.

L'enseignement de matières liées à la méthodologie d'enquête débute en deuxième année par un cours de *Théorie des sondages* de 48 heures (24 heures de cours et 24 heures de travaux pratiques). Ce cours se base sur les enseignements de *Statistique inférentielle* et de *Théorie des probabilités* acquis en première année. On y examine les spécificités de la théorie de l'échantillonnage, les techniques de planification, les méthodes d'estimation, les algorithmes de tirage, l'estimation de la précision, les problèmes liés aux erreurs de mesures, etc.

Pour les élèves attachés, cet enseignement est complété en deuxième année par un *Atelier de méthodologie d'enquête* où des professionnels issus principalement du secteur public exposent des problèmes méthodologiques et des études de cas réalisées dans leur propre milieu professionnel. Ensuite plusieurs modules de la formation continue diplômante des attachés sont consacrés à la méthodologie d'enquête.

En troisième année, seuls les élèves titulaires de la filière *statistique pour l'économie, les sciences sociales et la gestion* disposent d'un enseignement de

sondage. Dans cette filière qui regroupe depuis deux ans une quinzaine d'élèves, l'enseignement lié aux méthodes de sondage se décompose en trois parties :

1. Dans le cours de *Compléments de théorie des sondages*, on approfondit les connaissances théoriques et techniques dans les domaines de la théorie des sondages. Cette année, ce cours fut consacré aux problèmes d'estimation de précision, et plus particulièrement : au calcul de variance, aux techniques de linéarisation, aux erreurs de mesure, au traitement des non-réponses, etc.

2. L'*Atelier de méthodologie d'enquête* est une première ouverture vers la vie professionnelle. Les intervenants issus essentiellement du secteur privé y exposent des études de cas ou des questions relatives à leur pratique méthodologique dans leur environnement professionnel.

3. Le dernier enseignement intitulé *Réalisation d'une enquête statistique* est un travail collectif à quinze consistant à réaliser entièrement une enquête statistique. Au début de ce travail, les élèves ont donc suivi deux cours de sondage et un atelier. Ils sont donc censés maîtriser la théorie de l'échantillonnage et ont une idée des problèmes traités (et de leurs difficultés) dans les milieux professionnels. C'est donc l'occasion de mettre en pratique ce qui a été enseigné précédemment et surtout de réaliser un travail collectif.

Le projet "Réalisation d'une enquête statistique"

L'encadrement de la *Réalisation d'une enquête statistique* est assuré par trois enseignants : Patrick Lainé, Michel Simioni et Yves Tillé. Cet enseignement débute en novembre et doit se terminer à la fin du mois de mars, date de départ des élèves en stage. Les élèves disposent donc d'environ six mois pour réaliser entièrement l'enquête : s'approprier la thématique, définir le problème à traiter, construire un questionnaire, définir le plan de sondage, interroger les unités d'observations, construire une base de données, corriger la base, la traiter et rédiger un rapport.

Afin de donner au projet un aspect plus professionnel, un intervenant extérieur qui travaille dans le secteur public ou privé est sollicité pour jouer le rôle de commanditaire. Le "commanditaire" du projet n'est donc pas un statisticien mais un professionnel qui passe un contrat avec l'Ensai pour résoudre un problème auquel il est lui-même confronté. La convention passée entre le commanditaire et l'Ensai stipule que le travail est réalisé gratuitement par les élèves mais le commanditaire s'engage à couvrir les frais d'enquête qui peuvent être importants (de l'ordre de 32 000 francs pour l'enquête 1996-1997). Le choix du commanditaire est effectué par les trois enseignants en accord avec la direction de l'école.

Le commanditaire choisi pour la première enquête en 1996-1997 fut l'équipe d'Economie et de Sociologie Rurales de l'INRA de Rennes. Les élèves ont réalisé une enquête en face à face auprès de 700 ménages de la ville de Rennes sur la question du consentement à payer de consommateurs pour l'obtention d'une viande bovine exempte de risque (voir à ce sujet Abon et al., 1997, et Deletombe et al., 1998). L'enquête en cours de réalisation en 1997-1998 est commanditée par la Chambre Régionale de Commerce et d'Industrie de Bretagne. L'enquête porte sur l'évaluation des forces et handicaps de la sous-traitance bretonne par les donneurs d'ordres de cette région. Une trentaine d'entreprises ont été contactée pour un entretien en face à face avec le responsable des achats. Ensuite 300 entreprises seront contactées par téléphone sur cette question. Dans les deux cas, la méthodologie d'enquête a été définie complètement par les élèves de l'Ensai.

Il est évident que la principale difficulté de ce type d'enseignement consiste à organiser et à coordonner un travail en équipe. Le projet est coordonné lors d'une réunion hebdomadaire qui regroupe les élèves et les enseignants. Cette réunion est un lieu de discussion et de décision. Pour réaliser l'enquête, les élèves sont divisés en groupes. Un groupe assure la réalisation d'une tâche précise : appropriation de la thématique, réalisation du plan de sondage, rédaction du questionnaire, gestion des interviews, conception du masque de saisie, mise en place de la base de données, traitement des données, écriture du rapport. Chaque groupe d'élèves travaille sous la responsabilité d'un élève-coordinateur qui est l'interlocuteur privilégié des enseignants pour la réalisation de la tâche. Les groupes sont réorganisés au cours de l'avancement du projet et il est demandé à chaque élève de coordonner au moins une fois un groupe de travail.

L'évaluation se déroule en deux étapes : une étape individuelle et une étape collective. Le projet est d'abord soutenu collectivement par les élèves devant un jury composé d'un Président extérieur à l'école, du commanditaire et des enseignants. Ce jury donne une note globale aux élèves sur la base de leur travail. Les enseignants attribuent ensuite une note individuelle à chaque élève sur la base de son implication dans le projet. Cette note est attribuée après une discussion individuelle avec l'élève où l'on établit un bilan de ses activités dans le projet et où on fait état des difficultés et de l'intérêt de l'élève pour le projet. La note finale de chaque élève est la moyenne de la note collective et de la note individuelle.

Milieu scolaire et milieu professionnel

Une des ambiguïtés fondamentales de ce type d'enseignement qui vise à professionnaliser les savoirs scolaires est qu'une école n'est justement pas un environnement professionnel. La relation entre les enseignants et les élèves se distingue radicalement d'une relation professionnelle. La distinction porte essentiellement sur le partage des responsabilités dans la réalisation d'un travail.

Dans l'enseignement supérieur, on considère que les élèves sont responsables de leur réussite ou de leur échec. Même si l'on peut débattre longuement sur la responsabilité de l'échec scolaire, (débat dans lequel nous n'entrerons pas), dans l'enseignement supérieur, l'élève a le droit d'organiser lui-même son travail. Il en est donc individuellement responsable. Les enseignants interviennent très rarement dans l'organisation du travail des élèves. Quand des élèves sont soumis à un contrôle continu très "rapproché", ce qui est souvent le cas dans les Grandes Ecoles, ce suivi consiste avant tout à évaluer plus régulièrement les élèves et non à intervenir dans l'organisation de leur travail.

En milieu professionnel, l'organisation du travail est la responsabilité directe des directions. On reconnaît au directeur le droit d'organiser le travail et de distribuer les responsabilités. Le directeur est directement responsable de tout le travail de son département, principalement de la manière dont le travail est organisé. Cette notion de responsabilité ne peut se transposer directement à une Grande Ecole. Les enjeux ne sont pas les mêmes dans une école et dans une entreprise. Les élèves n'ont pas l'habitude d'exercer des responsabilités.

De plus, les élèves ont une "culture" à la fois très égalitaire, solidaire et individualiste. Aucun élève ne se mêlera de la manière dont s'organisent les autres. Même s'ils travaillent en binômes, ils se choisissent selon leurs affinités. En général, ce choix implique un accord sur la manière d'organiser le travail. Les élèves ne sont pas habitués à travailler ensemble. Quand ils ont des relations conflictuelles, ils ne doivent pas les gérer. Ils peuvent se contenter de ne pas communiquer entre eux. Ils ne demanderont jamais à un enseignant d'intervenir dans leurs querelles éventuelles. Une autre spécificité de la situation d'enseignant est que dans une école, l'élève a le droit à l'erreur. Ce droit est une condition nécessaire de l'apprentissage et est en contradiction avec les soucis de qualité exigée en milieu professionnel.

Difficultés organisationnelles

Nous pensons que la difficulté de ce type de démarche repose avant tout sur cette spécificité du milieu scolaire. Il est clair que l'organisation d'un projet de grande envergure (au sens où il implique une quinzaine d'élèves pendant plusieurs mois) amène nécessairement les enseignants à intervenir dans l'organisation même du travail. Rien ne serait plus démagogue que de laisser les élèves s'auto-organiser. Une telle démarche ne correspondrait d'ailleurs absolument pas à une mise en situation professionnelle. La réalisation d'un projet impliquant quinze élèves nous a amenés chaque année à considérer des problèmes qui d'ordinaire ne regardent pas les enseignants.

Une des composantes essentielles du travail professionnel est la notion de risque. En milieu professionnel, l'employeur organise le travail mais assume le risque. Dans

une école, l'enseignant ne gère pas le risque de l'échec de l'élève. Cependant, ce type de projet implique une gestion financière (plusieurs dizaines de milliers de francs). Un risque de sinistre n'est donc pas nul. De plus, le statut du Groupe des Ecoles Nationales d'Economie et de Statistique (Genes) rend difficile la gestion financière du projet : l'Ensai n'est pas habilité à gérer les flux financiers liés à un enseignement de ce type. En 1996-1997, l'argent du projet a ainsi été géré par le Bureau des Elèves. En 1997-1998, l'argent du projet a été géré directement par le commanditaire. Les élèves ont dû alors avancer des sommes parfois importantes sur leurs propres deniers.

La question du droit à l'erreur est également délicate. On peut raisonnablement penser qu'un projet de ce type peut échouer et échouera très probablement au moins une fois si l'expérience est répétée chaque année. En cas d'échec manifeste, il est impensable d'évaluer un groupe d'élèves très négativement au point de faire redoubler toute une promotion, ce qui rend l'évaluation collective assez caduque. Un échec important serait très gênant pour le commanditaire et dépasserait largement le cadre scolaire. La question du règlement d'un contentieux entre l'Ecole et le Commanditaire est à nouveau ambiguë puisque l'Ecole n'est financièrement pas partenaire de l'opération.

La mobilisation des élèves n'est d'ailleurs pas d'office acquise à ce type d'entreprise qui est à l'Ensai un enseignement obligatoire. De plus les élèves de troisième année sont souvent amenés à compléter leur formation en dehors de l'Ecole. Dans la filière *Statistique pour l'Economie et les Sciences Sociales et la Gestion*, un tiers des élèves sont inscrits dans un DEA. Enfin, ceux-ci sont généralement très préoccupés par leur entrée dans la vie professionnelle. Il leur faut trouver un lieu de stage. Ces divers éléments contribuent à rendre leur mobilisation assez hasardeuse en certaines périodes de l'année.

Gestion du projet

Malgré un certain nombre de difficultés, nous pensons que ce type d'enseignement peut s'avérer particulièrement intéressant. Dès le début, il suscite la perplexité des élèves et parfois l'enthousiasme. Ce n'est pas tellement l'aspect appliqué de l'enseignement qui rompt avec les habitudes car les élèves de l'Ensai ont déjà réalisé des applications statistiques dans le cadre des *projets statistiques* de première et deuxième année. L'aspect inhabituel du projet est surtout dû à son envergure : travail à 15, pendant plusieurs mois. La plus grosse difficulté consiste à organiser le travail en équipe dans un environnement qui ne s'y prête pas a priori. Après quelques semaines de travail, les élèves constatent leurs divergences concernant la manière de s'organiser. L'implication des élèves est également très variable. Dans certains cas, les difficultés entre élèves deviennent suffisamment gênantes pour que les enseignants s'en rendent compte clairement.

L'attitude des enseignants dans la gestion de ces difficultés ne peut être que très ambiguë. Si certains arbitrages peuvent être faits sur l'organisation du travail, il est difficile d'intervenir dans les relations des élèves. D'autant plus que leurs difficultés relationnelles prennent racines sur des questions parfois très personnelles. Dans ce contexte, nous pensons cependant que le rôle des enseignants doit se concevoir selon quatre axes :

1. Ressources et conseil

Le premier aspect de l'enseignement est le rôle de ressources et de conseils. Notre stratégie a consisté à laisser les élèves travailler seuls tant que nous considérions que tout se passait bien. Outre la réunion hebdomadaire, de nombreuses discussions, parfois dans les couloirs, permettent d'évaluer l'avancement du travail. L'essentiel est que les enseignants soient disponibles pour pouvoir efficacement dispenser quelques conseils.

2. Structuration du travail

Comme nous l'avons dit ci-dessus, nous pensons que les élèves partagent une culture foncièrement égalitaire. Les coordinateurs des groupes se trouvent donc dans un rôle tout à fait inhabituel et constatent qu'ils ne disposent d'aucun moyen coercitif pour obtenir quelque chose d'un de leurs pairs. Certains élèves bénéficient parfois d'un statut de leader naturel dans un groupe et sont capables d'organiser le travail. Cependant, nous demandons une rotation des coordinateurs et un changement dans la composition des groupes à chaque étape importante de l'élaboration du projet. Certains peuvent parfois se trouver dans des positions très embarrassantes. Ne disposant d'aucun moyen coercitif et se refusant à faire intervenir les enseignants, ils sont alors contraints de faire tout le travail du groupe tout seul. Dans ce contexte, la structuration du travail permet de prévenir ce type de problèmes. Cette structuration passe par une organisation du temps et l'établissement d'échéanciers.

3. Arbitrage

L'arbitrage en cas de difficultés est à notre avis nécessaire. Il est d'abord important de contrôler individuellement les élèves et de susciter une discussion en cas d'absentéisme flagrant. Parfois, les enseignants peuvent être amenés à intervenir dans des conflits et à préciser les responsabilités. Comme nous l'avons dit plus haut, ce type d'intervention n'est pas habituel en milieux scolaire ou universitaire mais il est parfois nécessaire pour permettre de trouver certaines solutions dans ce type d'enseignement et faire ainsi avancer la réalisation du projet.

4. Évaluation

Le dernier axe concerne évidemment l'évaluation des élèves, tâche ingrate que les enseignants ont trop tendance à négliger. L'évaluation est d'autant plus délicate qu'il est réellement difficile de discerner la contribution d'un élève lors d'un travail collectif. Outre l'évaluation par une soutenance collective du projet, nous avons opté pour une évaluation individuelle sur le mode professionnel. Lors d'un entretien individuel, nous faisons un bilan du travail de l'élève. Nous lui demandons de préciser les domaines sur lesquels il est intervenu, de donner un jugement sur son implication dans la réalisation du projet, de souligner les difficultés ou problèmes qu'il a pu rencontrer, d'émettre un avis sur l'organisation générale du projet et de faire, s'il le souhaite, des propositions d'amélioration. Ensuite, nous lui donnons notre opinion sur son travail en toute franchise. S'ensuit alors une discussion entre l'élève et les enseignants. Lors de la première évaluation en mars 1997, celle-ci a souvent porté sur l'apprentissage organisationnel tel qu'il a été perçu par les élèves.

Perspectives d'avenir

Il est clair que ce type d'enseignement n'est pas vraiment une innovation. Nous ne connaissons cependant que peu d'exemples d'enseignement de ce type donnés depuis plusieurs années dans un cadre scolaire. Il est pratiqué à Statistique Canada pour la formation de méthodologistes (voir Dumais, 1995) et, depuis deux ans, au Centre de Formation de l'Insee à Libourne (Cefil) pour la formation des contrôleurs de l'Insee. Mais il s'agit, dans ces deux cas, d'un enseignement donné en milieu professionnel. Si ce type de projet intéresse incontestablement les élèves, il pose de nombreuses questions dont la plus fondamentale est peut-être la suivante : comment inscrire un enseignement de ce type dans la durée ? Pour terminer, nous émettrons quelques suggestions à ce sujet.

L'implication et la disponibilité des enseignants sont une garantie de réussite. Cependant elle ne peut pas en être le seul élément. Etant donné les ambiguïtés de ce type d'enseignement, il nous semble important d'expliciter strictement les règles du jeu, les devoirs et les obligations de chacun. L'institution doit assumer clairement ses responsabilités en assumant le risque de l'expérience (moral et financier) avec le partenaire extérieur. Les enseignants doivent sortir de la relation habituelle qu'ils entretiennent avec des élèves en organisant réellement le travail et en s'organisant eux-mêmes pour être disponibles. Comme cela a été souligné ci-dessus, la partie la plus importante de l'organisation consiste pour eux à structurer le projet dans le temps : réunions hebdomadaires, échéancier, etc. Au cours de la réalisation du projet, ils sont aussi parfois amenés à arbitrer certains problèmes. Les conditions énoncées ci-dessus sont probablement nécessaires à une bonne organisation d'un projet mais ne garantissent nullement ni la motivation ni une bonne entente entre les élèves. La possibilité pour ceux-ci de choisir de s'engager ou pas dans la réalisation d'un tel projet, en accordant à cet enseignement un caractère optionnel, pourrait peut-être garantir la motivation des élèves le choisissant.

Abon T., Briand N., Chami, S., Chentouf, N., Clamens, M., Deletombe, O., Dubocage, E., Etchegoyen, M., Fillon, S., Gatignol, M., Gély, K., Huon de Pénanster, L., Jacquelain, V., Pastural, C., Rieg, C., et Yaacoub N. (1997). *Attitudes et consentements à payer des consommateurs pour obtenir des biens alimentaires à faible risque pour la santé : application à la maladie de la vache folle*, Rapport interne, Ensai, 163 pages.

Currie S.G. et al. (1986). « Preparing mathematical statisticians for statistical agencies », *Journal of Official Statistics*, 2.

Deletombe, O., Huon de Pénanster, L.,Lainé, P., Simioni, M., et Tillé, Y. (1998). « Enquête sur la sécurité alimentaire réalisée dans le cadre des enseignements de l'Ensai », à paraître dans : *Les sondages*, Université Rennes 2, 10 pages.

Dumais, J. (1997). « La formation des méthodologistes à Statistique Canada », dans : *Actes des Journées de Méthodalogie Statistique, 18 et 19 Octobre 1995*, Insee-Méthodes, n° 59-60-61, pages 47-67.

LE PROCESSUS DE RÉALISATION D'UNE ENQUÊTE PAR LES CONTROLEURS STAGIAIRES AU CEFIL

Bertrand Roucher

Le centre de formation de l'Insee à Libourne (Cefil) a été créé en 1996 .
Il a pour mission de mettre en œuvre :

- la formation des contrôleurs stagiaires
- des formations pour des stagiaires de pays étrangers
- des actions de formation continue pour le système statistique public

L'hébergement qui comporte une soixantaine de studios ou duplex est assuré pour tous les stagiaires à proximité du centre de formation. Ce dispositif joue un rôle important pour la cohésion des groupes.

Le cadre

Le dispositif de formation des contrôleurs stagiaires s'étend sur une année complète qui se décompose en trois périodes :

- 6 mois de formation à Libourne
- 3 mois de stage pratique
- 3 mois de phase d'adaptation à l'emploi

L'orientation de cette formation est résolument tournée vers l'apprentissage par la pratique en se basant sur le principe que l'on apprend vraiment ce que l'on réalise par soi-même. Naturellement, les aspects théoriques ou conceptuels sont abordés, soit en amont du travail s'ils sont nécessaires à sa réalisation, soit en aval lorsqu'ils viennent compléter ou formaliser la phase abordée.

Dans le cadre de la formation à Libourne, les contrôleurs stagiaires ont à réaliser, sur 5 semaines et demie, une enquête complète depuis la rencontre avec un commanditaire extérieur jusqu'à la présentation du travail et de ses résultats.

Les objectifs

Les objectifs de cette réalisation sont essentiellement pédagogiques. Mais s'il s'agit d'apprendre en réalisant, l'intérêt de l'opération est conditionné par l'appel à un commanditaire extérieur qui donne une motivation particulière à la réussite de l'opération.

Les objectifs reprennent les grandes phases de déroulement classique d'une enquête mais cette opération permet également de synthétiser et de mettre en oeuvre un grand nombre de connaissances et compétences déjà travaillées dans le déroulement du cursus.

Objectif 1 - Élaborer un système de recueil de données

Objectif 2 - Gérer efficacement une enquête

Objectif 3 - Produire des données normalisées après traitement de l'information

Objectif 4 - Restituer l'essentiel des informations contenues dans un ensemble de données et les mettre en forme

Objectif 5 - Rédiger une première analyse et la présenter oralement

Objectif 6 - Organiser et réaliser un travail en groupe

Les différents objectifs intègrent, pour chacun, les éléments suivants :

• **Objectif 1** ➔ Élaborer un système de recueil de données

- Réflexion sur la commande (la problématique)
- L'examen de l'existant
- Les différentes phases de déroulement du processus
- Les différentes modalités de collecte
- La conception de questionnaire
- Les modalités et les nomenclatures

• **Objectif 2** ➔ Gérer efficacement une enquête

- Le champ d'observation
- La source de l'information à la base de l'enquête, contrôle de qualité
- L'échantillonnage
- La mise en place, instructions, suivi, gestion des retours, rappels
- les contrôles manuels
- La construction d'un tableau de suivi

* **Objectif 3** ➔ Produire des données normalisées après traitement de l'information

- Les contrôles systématiques
- Les cohérences
- Le traitement des non-réponses
- Les redressements
- La représentativité
- Les recodifications
- Les tris à plat
- Le croisement de données

* **Objectif 4** ➔ Restituer l'essentiel des informations contenues dans un ensemble de données et les mettre en forme

- La sélection des informations pertinentes
- La mise en forme des tableaux et des graphiques
- Les règles de la déontologie statistique

* **Objectif 5** ➔ Rédiger une première analyse et la présenter oralement

- Les règles de construction de l'écrit
- L'intégration de tableaux et de graphiques dans le commentaire
- La présentation, les titres
- La restitution orale

* **Objectif 6** ➔ Organiser et réaliser un travail en groupe

- La gestion de projet
- L'organisation du travail, la répartition des tâches
- La communication dans un groupe

l'ensemble de ces objectifs mettent en oeuvre des compétences acquises au cours des sessions précédentes. Les stagiaires savent utiliser l'outil bureautique (traitement de texte, tableur). Ils ont été initiés au logiciel de traitement statistique (SAS), ils maîtrisent la statistique descriptive et ont assimilé les bases de la rédaction et de l'expression orale.

Le choix du sujet

La définition du thème de travail est une opération qu'il convient de préparer très en amont de la réalisation car, de sa réussite, dépend, pour une bonne partie, le succès final de l'opération.

Le partenaire doit être bien éclairé sur le rôle pédagogique de l'opération et accepter, explicitement, le risque de ne pas obtenir toutes les réponses à ses questions.

Le travail à effectuer doit répondre à un véritable besoin exprimé par le commanditaire. L'expression initiale de ce besoin est d'ailleurs d'une ampleur considérable. Il convient donc de canaliser la problématique proposée pour en extraire l'essentiel et la ramener à deux ou trois questions principales qui pourront être traitées dans le délai de cinq semaines.

Néanmoins, le commanditaire doit avoir un statut de service public, l'opération n'est pas facturée au partenaire. Celui-ci doit être de préférence implanté localement, car pour des raisons pratiques, nous souhaitons limiter le champ d'intervention du Cefil à la zone de Libourne.

Les bases utilisées pour l'enquête doivent être de bonne qualité et décrire correctement la population à observer.

Il convient également de déclarer l'enquête auprès de la Cnil en liaison avec le Département de la Coordination Statistique.

Enfin, de la problématique proposée, nous en extrayons deux ou trois sujets d'enquêtes complémentaires mais distinctes, afin de dégager ainsi une masse de travail suffisante pour plus de quarante stagiaires.

En 1997, le partenariat a été conclu avec la Mairie de Libourne. Celle-ci a pour objectif de maintenir l'attractivité commerciale du centre ville - que l'on notait en déclin - sans casser le dynamisme de la périphérie. De plus, la mairie a souhaité mesurer « l'évasion » commerciale (principalement sur Bordeaux), en connaître les causes et se donner les moyens de réduire le phénomène.

Cette problématique a fait l'objet de deux enquêtes simultanées décrivant les comportements d'achat des ménages (lieux de consommation, fréquence, nature des achats, moyens de transport, motivations et freins) :

- l'une auprès des ménages libournais
- l'autre auprès des consommateurs interrogés à la sortie des magasins

En 1998, le partenaire est la mission locale pour l'emploi et l'insertion des jeunes. Les trois axes de travail sont les suivants :

- Quelles sont les conditions de vie sociales et matérielles des jeunes inscrits à la Mission Locale et comment est constitué leur tissu familial et relationnel ?

• Quels ont été les parcours des jeunes de 22 à 25 ans, ayant fréquenté la Mission Locale, en quoi celle-ci leur a-t-elle apporté une aide, à quel moment ?
• Quelles sont les causes d'abandon du système scolaire, quelles sont les motivations et les projets des jeunes se trouvant dans cette situation ?

Si l'objectif de la Mission Locale réside d'abord dans une connaissance statistique de son propre public, elle souhaite retirer de ces travaux, des enseignements sur son approche des jeunes, sur les actions concrètes à proposer et - pourquoi pas - travailler en amont avec ses partenaires institutionnels.

L'organisation

En 1997, l'organisation du travail a été inspirée de l'expérience canadienne (intervention aux Journées Méthodologiques 1996 de J. DUMAIS).

Le travail de réalisation d'une enquête peut se découper en **4 grandes compétences** ayant en charge des tâches précises à effectuer : le terrain et la logistique, les méthodes, le traitement statistique et le traitement informatique. Ce découpage permet de répartir les stagiaires en groupes de travail chargés chacun d'une de ces 4 compétences.

- **Groupe 1 : terrain-logistique**

Les stagiaires ont en charge la communication interne et externe, le cahier des charges, la convention avec le partenaire, la gestion du budget de l'enquête, l'organisation matérielle de la collecte, le suivi général de l'enquête et enfin l'archivage des données de base.

- **Groupe 2 : méthodes**

Les stagiaires doivent commencer par réaliser une analyse des sources et de l'existant sur le thème de l'enquête, établir un profil démographique de la zone délimitée, définir le plan de sondage et le redressement des résultats, réaliser l'instruction de collecte et un plan de contrôle des réponses.

- **Groupe 3 : traitement statistique**

Les stagiaires ont en charge la rédaction du questionnaire, le test du questionnaire, le contrôle des questionnaires et le règlement des litiges et des non-réponses, la définition des tableaux en sortie et la rédaction du rapport final.

- **Groupe 4 : traitement informatique**

Les stagiaires commencent par se perfectionner aux outils informatiques qu'ils ont choisis. Ils ont en charge l'analyse informatique de la demande, le programme d'exploitation de l'enquête, l'édition des tableaux de résultats et la mise en forme du rapport final, incluant la réalisation de graphiques et des cartes, si nécessaire

En plus des travaux de groupe réalisés dans le cadre de compétences spécifiques, seront effectués :

- des apports auprès de l'ensemble de la promotion sous forme d'exposés ou de présentations-discussions,
- des tâches réparties entre tous les stagiaires, comme la collecte, la saisie des questionnaires, la première analyse des résultats et le rapport final,
- des réunions collectives pour valider certains choix, comme le questionnaire, l'analyse des résultats.

Le fonctionnement (en 1997)

Afin que chacun puisse participer activement à la réalisation des actions, chaque groupe de compétence a été constitué par 4 ou 5 personnes. Avec 41 stagiaires, on aboutit à la création de 8 groupes de 5 stagiaires. En conséquence, il s'est avéré nécessaire de traiter 2 angles de la problématique posée par le commanditaire et de réaliser ainsi 2 enquêtes de nature complémentaire.

Coordination : le projet de chaque enquête est suivi au cours d'une réunion quotidienne à laquelle participent les membres de l'équipe pédagogique et un représentant de chaque groupe de compétence. Ce « *comité de suivi* » a pour fonctions :

- de suivre l'état d'avancement du projet et notamment le suivi du planning
- d'étudier les problèmes qui se posent et de décider des solutions

Ces réunions se tiennent chaque jour. Le comité peut décider, en l'absence d'éléments nouveaux, de reporter sa réunion.

Organisation des groupes : chaque groupe de compétence a une série d'objectifs à atteindre et un planning à respecter dans le cadre de l'enquête à laquelle il participe. Le groupe est libre de choisir le mode d'organisation qui lui convient le mieux. Toutefois, il doit :

* ***désigner un représentant***, qui sera son porte-parole, lors de chaque réunion de coordination de l'ensemble du dispositif. Ce représentant établit un compte-rendu de ces réunions auprès des membres de son groupe.

* ***tenir à jour un cahier de bord*** précisant l'état d'avancement des travaux et les questions à résoudre. On y trouvera également les comptes-rendus des réunions de coordination.

Le groupe est aidé, en tant que de besoin, par un membre de l'équipe pédagogique qui joue le rôle de consultant. Ce dernier doit intervenir dans la vie ou l'organisation du groupe si des problèmes apparaissent.

Les apports

Si la réalisation de l'enquête nécessite des compétences déjà acquises au cours des sessions précédentes, elle a aussi nécessité - en cours de l'opération - des apports indispensables à son bon déroulement, notamment :

- la conception d'un projet d'enquête
- la conception d'un questionnaire
- l'initiation aux sondages
- en informatique : l'initiation à paradox pour la saisie
- le secret statistique

Ces apports ont été, en 1997, intégrés au cours de l'ensemble des cinq semaines et demi.

Le déroulement

Avec l'aide de l'équipe pédagogique, les stagiaires réalisent un planning de l'ensemble du dispositif dont les grandes étapes sont les suivantes :

1ère semaine

- Rencontre avec le commanditaire
- Détermination du planning
- Définition des objectifs
- Examen de l'existant
- Choix du mode de collecte
- Analyse de la base d'enquête
- Échantillonnage
- Mise au point d'un projet de questionnaire

2ème semaine

- Constitution d'une grille de saisie informatisée
- Test du questionnaire
- Relation avec la presse
- Repérage
- Début de collecte

3ème semaine

- Suite de la collecte
- Saisie des questionnaires
- Rappels
- Contrôles

4ème semaine

- Sortie des tableaux
- Analyse
- Rédaction

5ème semaine

- Fin de la rédaction
- Mise en forme
- Tirage
- Présentation des résultats au commanditaire

Le timing précis de ces opérations est délicat car certaines phases doivent souvent être enclenchées alors même que celle qui la conditionne n'est pas terminée (exemple : la constitution d'une grille de saisie informatisée est déjà largement amorcée alors que le questionnaire n'est pas encore définitif).

Le bilan

La collecte de l'enquête 1997 a permis de recueillir :

- 336 questionnaires pour les consommateurs interrogés à leur domicile,
- 505 questionnaires pour les consommateurs interrogés à la sortie du magasin.

La durée totale de la collecte n'a pas dépassé une semaine soit une moyenne globale de cinq questionnaires par jour et par stagiaire, le questionnement au domicile s'avérant moins productif que le questionnement sur la voie publique.

S'il convient de mesurer le bilan par confrontation aux objectifs initiaux, on peut considérer qu'ils ont été globalement atteints en 1997.

Le délai de cinq semaines et demie a été tenu, le questionnaire a été conçu, la collecte a été gérée et les traitements effectués. La présentation du travail au commanditaire a été réalisée au jour dit.

Les points les plus positifs ont concerné :

- l'organisation du projet

- le planning des groupes
- la collecte et la saisie
- la mise en oeuvre des compétences informatiques
- l'apprentissage - difficile - du travail en groupe
- le respect des délais

Les aspects les moins maîtrisés ont concerné :

- le repérage
- la qualité du questionnaire
- l'absence de test
- les instructions aux enquêteurs
- la tenue de tableaux de suivi de la collecte
- la coordination entre les différents groupes

➔ En 1997, l'organisation du travail a « spécialisé » chaque stagiaire sur quelques tâches mais ne lui a pas permis de partager la connaissance de l'ensemble des travaux. D'où un sentiment de frustration, pour quelques uns, à l'heure du bilan.

➔ La difficulté à répartir correctement la charge de travail sur l'ensemble de la séquence a également entraîné de fortes surcharges de travail, en certaines périodes alors que d'autres ont été nettement plus calmes.

➔ Le positionnement de l'équipe pédagogique a été jugé, par les stagiaires, un peu trop distancié. Du point de vue de l'équipe pédagogique, il s'agissait de responsabiliser totalement les groupes mais ceux-ci ont eu le sentiment d'avoir à trop se débrouiller seuls.

➔ La tenue du « cahier de bord » a laissé à désirer car l'investissement a été inégal selon les groupes.

Les modifications apportées au dispositif en 1998

Le déroulement général de l'opération est inchangé et malgré les difficultés, la durée de cinq semaines et demie pour la réalisation de l'enquête a été maintenue. Les principales modifications concernent les apports et surtout l'organisation.

• Les apports (sondage, secret statistique) ont été traités en amont de la séquence, de façon à dégager un temps supplémentaire pour l'opération. De plus, une enquêtrice de la Direction Régionale d'Aquitaine intervient cette année sur les aspects pratiques de présentation et de comportement face à l'enquêté.

- La répartition des stagiaires par groupes de compétences a été abandonnée. Chaque groupe de quinze stagiaires est responsable en totalité de l'ensemble d'un axe de travail et donc d'une enquête spécifique. Chaque groupe doit assurer le partage des compétences et chacun pourra ainsi réaliser sa part de travaux en fonction de ses préférences dans le dispositif, tout en conservant la visibilité sur l'ensemble du projet.

- Un membre de l'équipe pédagogique suit particulièrement chaque groupe et joue à son égard le rôle de *maître d'ouvrage*.

- Un journal quotidien est réalisé et installé sur le site Intranet expérimental du Cefil. La Division de la Formation, les contrôleurs de la promotion précédente et les Responsables de Formation ont eu communication du code permettant l'accès au serveur du Cefil.

Conclusion

Cette opération de réalisation d'enquête est assez difficile à monter : choix du partenaire, détermination du sujet, nombre de stagiaires, organisation du dispositif... Les écueils ne manquent pas et les conditions de réalisation sont encore extrêmement aggravées par la durée que nous avons souhaité y consacrer, durée improbable dans un processus normal.

Néanmoins, la première expérience, sans être une réussite parfaite, a montré que les objectifs initiaux pouvaient être atteints. La seconde expérience, vécue actuellement, confirme largement que la réussite sera encore au bout.

Bien entendu, les stagiaires ont investi (et investissent), sans compter leur temps, pour répondre au challenge proposé. Le fait de présenter un travail avec une finalité concrète, de mener à bien l'opération, malgré toutes les contraintes, donne aux stagiaires, en plus de compétences pratiques ou théoriques, un sentiment de confiance quant à leur capacité d'organisation et de mobilisation.

Mais, pourra-t-on trouver tous les ans un commanditaire compréhensif, un sujet digne d'intérêt et des stagiaires très motivés ? C'est une autre question dont nous pensons que la réponse doit rester positive, le plus longtemps possible.

L'ENSEIGNEMENT DE LA PRATIQUE DES ENQUÊTES À L'ENSEA (ABIDJAN, CÔTE D'IVOIRE)

Benjamin Zanou

L'Ecole Nationale Supérieure de Statistique et d'Economie Appliquée (Ensea) a pour vocation de former des statisticiens pour les pays africains d'expression française.

Depuis sa création en 1961, elle s'est développée progressivement en s'adaptant aux réalités socio-économiques de la sous-région et aux besoins en statistiques diverses.

Aujourd'hui, l'Ecole assure la formation des statisticiens au sein de quatre filières en fonction du niveau de recrutement des élèves :

- Ingénieurs Statisticiens-Economistes (ISE) en 3 ans ;
- Ingénieurs des Travaux Statistiques (ITS) en 2 ans (en 4 ans jusqu'à la promotion 98) ;
- Adjoints Techniques de la Statistique (AD) en 2 ans ;
- Agents Techniques de la Statistique (AT) en 1 an.

L'effectif des élèves des quatre filières tourne autour de deux cents (200). L'Ensea propose par ailleurs, des actions d'initiation et de perfectionnement dans différentes matières, destinées aux cadres de l'administration publique et privée.

Parallèlement à cette vocation de formation, l'Ecole développe progressivement, ses capacités en matière d'études et de recherche. Dans ce cadre, elle s'est dotée d'un département de recherche dont les objectifs sont :

- initier très tôt les élèves à la recherche en les associant à des activités de collecte sur le terrain, d'exploitation informatique et d'analyse des données ;
- réaliser des études et travaux de recherche de haut niveau, notamment en collaboration avec des centres ou instituts de recherche nationaux et étrangers.

En dehors des cours théoriques dispensés dans les matières classiques et les statistiques appliquées, l'Ecole a institué un cours de pratique des enquêtes dispensés dans les filières ISE, ITS et AD. Ce cours comporte deux parties : une partie théorique et une autre pratique.

L'organisation du cours à l'Ensea

Dans sa conception, le cours de la pratique des enquêtes est organisé de la même façon dans les trois filières ; mais il dépend surtout du nombre d'heures imparti à chacune d'elle.

1.1 - L'organisation du cours en division Adjoint Technique

Le volume horaire de ce cours est de 40 heures réparties en 20 heures de cours et 20 heures de travaux pratiques. Il est destiné aux AD de la deuxième année. Le professeur consacre les 20 premières heures aux éléments d'une enquête et les 20 autres heures aux exposés des élèves sur des enquêtes déjà réalisées. Ces exposés portent sur la méthodologie, le questionnaire, le plan de sondage, etc.

En dehors du cours, les élèves Adjoints Techniques participent à l'exécution de l'enquête à but pédagogique conçue par l'Ecole. Ils sont utilisés sur le terrain comme Agents enquêteurs et participent à la codification et à la saisie des données collectées en tant que stagiaires.

1.2 - L'organisation du cours en division Ingénieurs des Travaux Statistiques

Compte tenu du nombre d'heures qui lui est consacré *(40 heures initialement et 50 heures depuis deux ans),* le cours de la pratique des enquêtes bénéficie plus aux ITS qu'aux deux autres divisions. En effet, avec 50 heures en première année et 20 heures *en 2e année*, les élèves reçoivent le cours théorique, participent à la conception d'une opération de collecte, réalisent la collecte, exploitent les données et analyse les résultats.

1.2.1- L'organisation du cours en première année

En une vingtaine d'heures, le professeur expose les différentes phases d'une enquête en insistant sur les difficultés à chaque étape.

A la suite de cette phase, il est demandé aux élèves de proposer des thèmes d'enquête au cas où la direction de l'Ecole n'en a pas *prévu. Dans ce cas,* les discussions, alimentées par les arguments des uns et des autres amènent à retenir l'une des propositions comme le thème de l'enquête de l'année.

A partir de cet instant, des groupes de travail de 4 à 5 élèves sont formés au niveau de la classe. Ainsi, commence la préparation de l'enquête. Chaque phase de la préparation est abordée en séance de classe (définition des objectifs, préparation des correspondances, élaboration du calendrier, confection du questionnaire, rédaction des manuels d'instruction aux enquêteurs, élaboration des divers bordereaux, etc.).

Chaque séance de préparation est dirigée par un groupe avec un président de séance et deux rapporteurs chargés de faire le compte rendu de la séance. Le professeur n'intervient que pour apporter un éclaircissement et parfois pour départager les *protagonistes.* Les groupes sont chargés des tâches ponctuelles suivant l'évolution des activités préparatoires. Ainsi, il peut être demandé à un groupe de faire une proposition de maquette du questionnaire ou rédiger le manuel d'instructions aux enquêteurs sur les variables 1 à 10, etc.

Au cas où le thème de l'enquête est proposé par la direction de l'école, compte tenu des intérêts du bailleur de fonds, la préparation se fait de la même façon, sauf que le professeur oriente plus les débats dans le sens de ce qui est attendu. De toutes façons, les ITS4 stagiaires travaillent d'abord sur le sujet avant de soumettre les résultats à discussion.

Lorsque tous les documents sont prêts, les élèves exécutent l'enquête sur le terrain, soit sous la seule supervision du professeur du cours de Pratique des Enquêtes quand l'enquête ne concerne que les ITS3, soit sous la supervision des professeurs permanents de l'Ecole. Pour l'exécution de l'enquête sur le terrain, des élèves plus expérimentés (ceux qui ont déjà participé à ce genre de travail soit au cours de leur scolarité, soit pendant une vie professionnelle antérieure) sont choisis comme contrôleurs lors de la collecte.

1.2.2- Les retombées de l'enquête en ITS2

Après avoir suivi le cours et participé à la collecte en ITS1, les élèves de la deuxième année font l'exploitation et l'analyse de données d'enquête à travers le cours d'analyse des données. Il peut s'agir ou non d'une partie des données de l'enquête réalisée en ITS1.

1.3 - L'organisation du cours en division Ingénieurs Statisticiens-Economistes

A ce niveau, l'objectif du cours est de donner un aperçu des difficultés rencontrées dans l'élaboration et l'exécution d'une enquête. Les 20 heures sont consacrées à l'examen des différentes phases de la conception et de l'exécution d'une opération de collecte. Les élèves sont associés ensuite à la collecte sur le terrain *comme agents enquêteurs.*

II - L'organisation des opérations de collecte à l'Ensea

Depuis le début des années 1980, le Fonds des Nations Unies pour la Population (FNUAP) assiste l'Ensea pour la formation des Ingénieurs des Travaux Statistiques orientés vers les activités de population. Ce programme contient un volet recherche sur les questions de population. C'est celui-ci qui permet d'organiser tous les ans, une opération de collecte à l'intention des élèves.

Comme il est mentionné plus haut, le thème de l'enquête est choisi soit par le corps professoral, soit par les étudiants eux-mêmes. Dans un cas comme dans l'autre, il est demandé à deux élèves Ingénieurs des Travaux Statistiques en fin de cycle d'effectuer leur stage à partir du mois de février sur *l'enquête.* Ils deviennent responsables de l'étude sous la direction du professeur chargé du cours. Ce sont eux qui finalisent les documents discutés avec l'ensemble des ITS3 (dans l'ancienne formule) ou des ITS1 (dans la nouvelle formule). Ils supervisent la collecte et s'occupent après le terrain de l'exploitation et de la rédaction du premier rapport.

Lorsque les outils de collecte sont prêts et la méthodologie de l'opération arrêtée, une enquête pilote est organisée. Elle voit la participation des élèves intéressés (AT, AD2, ITS1, ISE2), et se déroule en une journée dans une localité proche de la ville d'Abidjan. Cette opération qui est le premier contact des élèves avec le terrain permet de tester le questionnaire et la méthodologie mais aussi la réaction des élèves face au terrain. Après l'enquête pilote, des séances de classe sont organisées avec les étudiants pour tirer les enseignements et finaliser les documents et la méthodologie.

La période de la collecte proprement dite se situe généralement vers la fin du mois de mars et le début du mois d'avril.

A cette période, toutes les sections intéressées par l'enquête vont sur le terrain pour deux semaines. Ce départ est précédé d'une formation de deux à trois jours sur le questionnaire de l'enquête. Les élèves sont regroupés en équipes de 7 à 10 personnes avec un chef d'équipe qui généralement a déjà une expérience d'enquête.

L'ensemble des équipes est encadré par le corps professoral en particulier le professeur responsable de l'enquête.

Pendant la collecte, des séances de travail sont organisées tous les soirs pendant les quatre premiers jours et un soir sur deux le reste du temps. Ces séances se font avec l'ensemble des étudiants et permettent de faire le point sur l'avancement des travaux et proposer des solutions aux difficultés rencontrées sur le terrain.

Au retour du terrain, les élèves ITS1, AD2 et AT s'occupent de la codification et éventuellement de la saisie des données. Les programmes de saisie et de l'exploitation étant conçus par les Ingénieurs des Travaux Statistiques qui font leur stage sur *l'enquête avec l'aide de personnes plus expérimentées (anciens élèves, professeurs de l'école).* L'analyse des premiers résultats est assurée également par eux.

III - Problèmes soulevés par l'organisation de cet enseignement

L'objectif final du cours étant de faire acquérir aux élèves *une première* pratique des enquêtes, il est indispensable de les faire participer aux différentes phases de l'organisation et de l'exécution d'une telle opération . c'est à ce niveau que se pose quelques problèmes.

3.1 - Le financement

L'ambition de la direction de l'Ensea a toujours été de faire participer tous les élèves en même temps à l'opération de collecte. Mais cela n'a jamais pu se réaliser pour une question de moyens financiers. En effet, pour utiliser deux cents enquêteurs pendant deux semaines, il faut des moyens financiers que l'Ensea n'a jamais pu réunir pour de telles opérations.

3.2 - Le recours à du personnel extérieur

En dehors des élèves et des professeurs permanents, il arrive parfois que la direction de l'Ensea fasse appel à des agents de l'Institut National de la Statistique (INS) ou recrute des agents temporaires.

L'appui de l'INS est demandé surtout pour la mise à jour cartographique des localités d'enquête afin de bien identifier les unités géographiques du premier et du second ordre pour éviter aux agents d'empiéter sur une autre unité ou d'omettre une

partie de leur zone. Il arrive également qu'un cadre de l'INS soit sollicité pour intervenir dans l'encadrement sur le terrain.

Pour certaines études portant sur la santé de la reproduction, compte tenu de la délicatesse de certaines questions adressées aux femmes (date des premières règles, reprise des rapports sexuels, etc.), il est procédé au recrutement des enquêtrices expérimentées pour administrer les questionnaires femme et homme tandis que le questionnaire-ménage est rempli par les étudiants.

3.3- L'encadrement

Il est vrai que la direction de l'Ecole accorde une grande importance aux cours théoriques de la pratique des enquêtes ainsi qu'à l'enquête sur le terrain. Mais, ces travaux de terrain ne sont pas prévus dans le programme annuel des étudiants. Les cours habituels n'étant suspendus que pour les élèves impliqués dans la collecte des données, les professeurs ne sont pas totalement dégagés des enseignements pour assurer l'encadrement de ces élèves sur le terrain.

L'encadrement des élèves sur le terrain est assuré principalement par le professeur responsable de l'enquête. Il est assisté par deux ou trois professeurs permanents qui parviennent à aménager leur emploi du temps. Même si ceux-ci ne disposent pas toujours d'une expérience dans le domaine des enquêtes, leur présence sur le terrain est toujours appréciée par les élèves.

De plus, le volume horaire pour le cours de la pratique des enquêtes n'étant pas le même d'une division à une autre, tous les étudiants (en particulier les ITSS et les ISE) ne reçoivent pas la même formation.

3.4- Evaluation des élèves

Certains élèves sont évalués par rapport au cours de la pratique des enquêtes et d'autres non.

Les ISE2 reçoivent une note qui est la moyenne de deux autres :

- une note d'un examen écrit portant sur le cours théorique de la pratique des enquêtes ;

- une note qui est l'évaluation de la pratique. Elle est basée sur l'assiduité au cours, la participation à la préparation de l'enquête, le travail sur le terrain et le comportement sur le terrain.

- Les ITS1 sont évalués uniquement sur la base des travaux de préparation et d'exécution de l'enquête avec les mêmes critères que les ISE2.

- Quant aux AD2 et AT, ils ne sont évalués sur l'enquête que si elle leur sert de sujet de stage. Dans ces conditions, les critères d'évaluation sont les mêmes que ceux des ISE2 et ITS1.

3.5 - La perception des Enquêtes Ensea par la population

Avant d'entreprendre toute opération sur le terrain, l'Ensea par l'intermédiaire du Ministre de tutelle adresse une demande d'autorisation au Ministre de l'Intérieur. Par cette correspondance, elle sollicite le soutien des autorités préfectorales, sous-préfectorales, communales et villageoises pour le bon déroulement de l'enquête. Une fois sensibilisée, la population se prête facilement aux questions des enquêteurs.

IV - Le point de vue des étudiants

Les élèves assistent au cours théorique de la pratique des enquêtes comme aux autres cours (certains sont intéressés et d'autres moins). L'intérêt *s'amplifie* quand il s'agit de préparer une opération de collecte. Les séances de classe sont très animées et parfois houleuses.

Certains élèves ont quelques appréhensions quand il s'agit de partir sur le terrain : ils posent des questions sur les conditions d'hébergement, de restauration, et d'hygiène. Mais au bout de deux semaines de terrain, on assiste à la situation inverse où certains *d'entre eux* souhaiteraient poursuivre la collecte pour avoir trouvé les conditions agréables et pour s'être faits des amis. Ils se plaisent à dire qu'ils souhaitent renouveler l'expérience et qu'ils ont beaucoup appris d'autant qu'il y a une différence importante entre les réalités du terrain et les prévisions du bureau.

Les étudiants sont satisfaits dans l'ensemble de ce cours même s'ils déplorent le fait qu'ils ne maîtrisent pas toutes les phases de la conception et de l'exécution d'une enquête. Il est évident qu'un cours de 20 à 50 heures ne permet pas de maîtriser toutes les techniques d'enquête et l'on sait par expérience que c'est à force de mener des enquêtes qu'on devient un spécialiste.

Pour notre part, il est apparu à travers les échos qui nous parviennent que les élèves qui ont participé à la préparation et à l'exécution de l'une de nos enquêtes sont à même de contribuer activement à des opérations analogues surtout s'ils ont effectué leur stage de fin de cycle dans ce cadre.

- Publications issues des enquêtes Ensea
- Récapitulatif des enquêtes effectuées à l'Ensea de 1982 à 1997
- Calendrier de l'Enquête Démographique et de Santé dans la sous-préfecture de Niakaramandougou - 1997.

ADER YA Kouadio Etienne et al (1982)
Recensement des villages de Memni et de Montézo ; Etudes et Recherches de l'Ensea ; pp-59.

KONE Harouna et al (1984)
Enquête sur l'activité économique et l'habitat des villages de Memni et de Montézo - Méthodologie - éléments d'analyse ; Etudes et Recherches de l'Ensea ; pp 28.

KOFFI N'Guessan et Benjamin ZANOU (1985)
La population de la commune de Jacqueville, recensement d'Avril 1984 ; Etudes et Recherches de l'Ensea ; pp.61.

ADOU Ayékoué et KOFFI N'Guessan (1987)
La population de Brobo : analyses socio-démographiques à partir d'une enquête par sondage réalisée auprès de 20 villages de la sous-préfecture de Brobo. Département de Bouaké ; Etudes et Recherches de l'Ensea ; pp.40.

Patrice VIMARD (1987)
Structure des ménages en pays Baoulé : composition et typologies familiales à Brobo ; Etudes et Recherches de l'Ensea ; pp.40.

COULIBALY Drissa et al (1987)
La population de la commune de Boundiali - Recensement d'Avril 1987 ; Etudes et Recherches de l'Ensea ; pp.142.

ANDRIAMAMPAHERY Dimby et al (1989)
Dynamique de population et mutation économique dans le Sud-Ouest ivoirien : la sous-préfecture de Sassandra ; Etudes et Recherches de l'Ensea ; pp.143.

KOFFI N'Guessan, Benjamin ZANOU (1989)
Analyse socio-économique et démographique d'une commune de la boucle du cacao - Daoukro : la population ; Etudes et Recherches de l'Ensea ; pp.153.

EDI Serge Jean (1989)
Analyse socio-économique et démographique d'une commune de la boucle du cacao - Daoukro : Les entreprises et leurs caractéristiques ; Etudes et Recherches de l'Ensea pp.52

La population de la commune de Biankouma (Ouest de la Côte d'Ivoire), Avril 1991.

KOFFI N'Guessan et al (1998)
Étude socio-Démographique et de planification familiale dans la commune de Tanda ; Etudes et Recherches de l'Ensea ; pp.100

ZANOU Benjamin et al (à paraître)
Fécondité et santé de la reproduction à Memni et Montézo. Etudes et Recherches de l'Ensea .; pp.61.

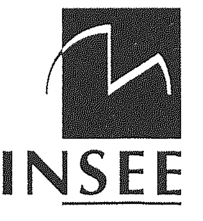
ZANOU Benjamin et al (1998)
Enquête démographique et de santé dans la sous-préfecture de Niakaramandougou . Études et Recherches de l'Ensea ; pp.93.

Récapitulatif des enquêtes effectuées à l'Ensea avec les étudiants de 1982 à 1997

Thème de l'enquête	Date et partenaire	Observation
Recensement des villages de MEMNI et de MONTEZO	Février 1982 FNUAP	Il s'agit du recensement des 2 villages avec quelques caractéristiques socio-démographiques et économiques de la population. Ce sont 2 villages de 7 800 habitants situés à 70 km d'Abidjan.
Enquête sur l'activité économique et l'habitat des villages de Memni et Montézo	Avril 1983 FNUAP	Au recensement exhaustif de la population a été greffée une enquête sur l'activité économique de la population et l'habitat.
Recensement de la population de la commune de Jacqueville	Avril 1984 FNUAP	Jacqueville est située sur le littoral à 65 km d'Abidjan. Un recensement exhaustif avec quelques caractéristiques de la population y a été effectué.
Enquête socio-démographique de Brobo	Août 1986 UNICEF	Il s'agit d'une enquête par sondage auprès de 20 villages au centre du pays (près de 400 km d'Abidjan).
Recensement de la population de la commune de Boundiali	Avril 1987 FNUAP	La commune de Boundiali est située à plus de 700 km d'Abidjan au Nord de la Côte d'Ivoire. Elle comptait 22 500 habitants environ.
Dynamique de population et mutation économique dans le Sud-Ouest ivoirien : La sous-préfecture de Sassandra	Avril 1988 FNUAP	Zone de peuplement récent, la sous-préfecture de Sassandra fait partie du nouveau front pionnier de la Côte d'Ivoire pour la culture du couple café-cacao. Autochtones, allochtones et étrangers se côtoient. L'objectif de l'étude est de mesurer cette mutation économique. Sassandra est située sur le littoral à 300 km environ d'Abidjan.
Etude Socio-économique et démographique d'une commune de la boucle du cacao : Daoukro	Avril 1989 FNUAP	Située à 250 km environ d'Abidjan, la commune de Daoukro fait partie de l'ancienne boucle du cacao. l'Enquête vise à mesurer le niveau économique après le déplacement de la boucle du cacao.
Fréquentation des « maquis » du campus universitaire par les étudiants.	Avril 1990 Ensea	Seuls les élèves ITS3 ont réalisé cette enquête qui n'a donné lieu à aucun rapport d'analyse.
La planification familiale dans la commune de Jacqueville	Avril 1993 FNUAP	Neuf ans après la première opération, Jacqueville a fait l'objet d'une enquête CAP sur la planification familiale. Le rapport n'est pas encore publié.
Etude Socio-démographique et de planification familiale dans la commune de Tanda.	Avril 1994 FNUAP	Tanda est la première commune du Nord-Est touchée par les études de l'Ensea. Peuplée de 18 000 habitants, la commune à été recensée entièrement, mais l'enquête n'a porté que sur le centre urbain.
Enquête sur la perception de l'utilisation des préservatifs en milieu scolaire et universitaire à Abidjan	Mai 1996 Ensea	Etude réalisée par les ITS3 uniquement. Aucun rapport n'est pour le moment rédigé. Elle s'est déroulée dans quelques établissements secondaires et universitaires. La taille de l'échantillon est de 1 200.
Fécondité et santé de la reproduction à Memni et Montézo	Mars 1997 Coopération Française	Quatorze ans après la dernière opération l'Ensea est repassée à Memni et Montézo pour mesurer la dynamique de la population avec un accent sur la santé de la reproduction. La population est de 10 500 habitants.
Enquête démographique et de santé dans la sous-préfecture de Niakaramandougou	Mars 1997 FNUAP Coût : 9 832 000	Après les observatoires de Memni et Montézo, Aboisso et Sassandra, l'Ensea a ouvert en 1997 un observatoire en zone de savane. Après le recensement de la population de dix localités dont le chef-lieu de la sous-préfecture, des questions ont été posées à toute la population sur sa santé au cours des trois mois précédant l'enquête. Population : 15 000 habitants.

Calendrier définitif de l'Enquête Démographique et de Santé dans la sous-préfecture de Niakaramandougou - 1997

ACTIVITÉ	DÉBUT	FIN
1. 1ère visite	22/01/97	22/01/97
2. Sensibilisation	22/01/97	31/03/97
3. Établissement d'un calendrier	03/02/97	03/02/97
4. Élaboration des documents administratifs	03/02/97	15/02/97
5. Recherche documentaire	03/02/97	31/05/97
6. Recherche cartographique	04/02/97	24/02/97
7. Élaboration du questionnaire et document de base	10/02/97	24/02/97
8. Acquisition du matériel	26/02/97	03/03/97
9. Enquête pilote	01/03/97	01/03/97
10. Finalisation des documents	03/03/97	08/03/97
11. Impression des documents	10/03/97	15/03/97
12. Formation des agents	17/03/97	18/03/97
13. Collecte des données	19/03/97	29/03/97
14. 2ème visite dans la zone	06/03/97	09/03/97
15. Dépouillement manuel	03/04/97	08/04/97
16. Codification	07/04/97	03/05/97
17. Saisie	28/04/97	05/05/97
18. Apurement des fichiers et tabulation	15/04/97	15/05/97
19. Analyse	16/05/97	30/11/97
20. Impression du rapport d'analyse	01/12/97	15/12/97
21. Diffusion du rapport d'analyse	16/12/97	31/12/97



L'INFORMATION SUR L'INFORMATION

INSEE ACTUALITES

"INSEE ACTUALITÉS magazine" est un catalogue trimestriel des nouveautés de l'INSEE : publications, banques de données... ; il est adressé à toute personne ou organisme désireux de suivre l'actualité de l'INSEE.

Abonnement gratuit sur simple demande à :
Insee - Direction générale
Abonnement à Insee Actualités - Timbre H533
18 bd A. Pinard - 75675 Paris cedex 14

COURRIER DES STATISTIQUES

Quatre fois par an cette revue interministérielle vous informe sur l'ensemble des activités du système statistique public et sur l'évolution des outils et des méthodes.

Abonnement 1 an (4 numéros)
France : 135 FF - Europe : 169 FF - Reste du monde : 234 FF
20,58 euros 25,76 euros 35,67 euros

LES PÉRIODIQUES

LE BULLETIN MENSUEL DE STATISTIQUE

10 000 séries mensuelles, trimestrielles et annuelles concernant l'ensemble de la vie économique, complétées par les séries rétrospectives des principaux indices et par le bilan démographique.

Abonnement 1 an (12 numéros)
France : 364 FF - Europe : 455 FF - Reste du monde : 584 FF
55,49 euros 69,36 euros 89,03 euros

ÉCONOMIE ET STATISTIQUE

Chaque numéro est un recueil d'articles sur un grand thème du débat social proposant des commentaires, des tableaux et des graphiques ainsi qu'une bibliographie.

Abonnement 1 an (10 numéros)
France : 414 FF - Europe : 518 FF - Reste du monde : 633 FF
63,11 euros 78,97 euros 96,50 euros

ANNALES D'ÉCONOMIE ET DE STATISTIQUE

Ce trimestriel publie des travaux originaux de recherche théorique ou appliquée dans les domaines de l'économie, de l'économétrie et de la statistique.

Abonnement 1 an (4 numéros)
France : 517 FF - Europe : 646 FF - Reste du monde : 691 FF
78,82 euros 98,48 euros 105,34 euros
Pour les particuliers :
France : 188 FF - Europe : 235 FF - Reste du monde : 278 FF
28,66 euros 35,83 euros 42,38 euros

INSEE MÉTHODES

La méthodologie des travaux de l'INSEE et les modèles.

Abonnement (10 numéros)
France : 728 FF - Europe : 910 FF - Reste du monde : 1 103 FF
110,98 euros 138,73 euros 168,15 euros

INSEE PREMIÈRE

Le "4 pages" qui, chaque semaine, présente les analyses et les commentaires des experts de l'INSEE sur un thème de l'actualité économique et sociale.

Abonnement (60 numéros)
France : 530 FF - Europe : 663 FF - Reste du monde : 827 FF
80,80 euros 101,07 euros 126,08 euros

INSEE RÉSULTATS

Cette série présente les résultats détaillés des enquêtes et opérations statistiques menées par l'INSEE.
Elle s'articule en 5 thèmes :

Économie générale (20 numéros)
France : 1 454 FF - Europe : 1 818 FF - Reste du monde : 2 075 FF
221,66 euros 277,15 euros 316,33 euros

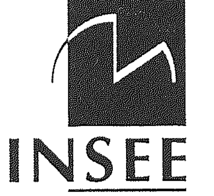
Démographie - Société (7 numéros)
France : 509 FF - Europe : 636 FF - Reste du monde : 726 FF
77,60 euros 96,96 euros 110,68 euros

Consommation - Modes de vie (10 numéros)
France : 728 FF - Europe : 910 FF - Reste du monde : 1 050 FF
110,98 euros 138,73 euros 160,07 euros

Système productif (15 numéros)
France : 1 091 FF - Europe : 1 364 FF - Reste du monde : 1 557 FF
166,32 euros 207,94 euros 237,36 euros

Emploi - Revenus (18 numéros)
France : 1 308 FF - Europe : 1 635 FF - Reste du monde : 1 860 FF
199,40 euros 249,25 euros 283,56 euros

Ensemble des 5 thèmes (70 numéros)
France : 5 090 FF - Europe : 6 363 FF - Reste du monde : 7 259 FF
775,97 euros 970,03 euros 1 106,63 euros



COLLECTION SYSTÈME STATISTIQUE PUBLIC

RECUEIL D'ÉTUDES SOCIALES

Une sélection d'études sur l'actualité sociale les plus récemment publiées par différents organismes publics français d'études et de statistiques.

Abonnement 1 an (3 numéros)
France : 315 FF - Europe : 394 FF - Reste du monde : 421 FF
48,02 euros 60,06 euros 64,18 euros

SYNTHÈSES

Cette nouvelle collection présente des études et des enquêtes faites par les organismes du système statistique public.

Abonnement 1 an (6 numéros)
France : 436 FF - Europe : 545 FF - Reste du monde : 623 FF
66,47 euros 83,08 euros 94,98 euros

LA CONJONCTURE COLLECTION "INSEE CONJONCTURE"

INFORMATIONS RAPIDES

Série de 350 numéros par an, présentant dès leur disponibilité les derniers indices et les résultats les plus récents des enquêtes de conjoncture de l'INSEE.

Elle inclut les 105 numéros des "Principaux indicateurs" (chiffres essentiels de l'économie) qui peuvent faire l'objet d'un abonnement à part par courrier ou par télécopie.

Abonnement
Principaux Indicateurs (105 numéros par an):
. par télécopie :
France : 2 000 FF - Europe : 2 500 FF - Reste du monde :3 000 FF
304,90 euros 381,12 euros 457,35 euros
. par courrier :
France : 830 FF - Europe : 1 038 FF - Reste du monde :1 288 FF
126,53 euros 158,24 euros 196,35 euros

Abonnement Informations Rapides (245 numéros par courrier) + les principaux Indicateurs par télécopie :
France : 2 950 FF - Europe : 3 688 FF - Reste du monde :4 563 FF
449,72 euros 562,23 euros 695,62 euros
Abonnement à l'ensemble de la série par courrier :
France : 1 750 FF - Europe : 2 188 FF - Reste du monde :2 691 FF
266,79 euros 333,56 euros 410,24 euros

TABLEAU DE BORD HEBDOMADAIRE

Un panorama complet et actualisé de la conjoncture économique française et internationale. Le supplément "Série longues " donne des tableaux et des graphiques sur 25 ans.

Abonnement 1 an (50 numéros + 1 supplément Séries longues)
France : 1 500 FF - Europe :1 875 FF - Reste du monde : 2 375 FF
228,67 euros 285,84 euros 362,07 euros

NOTE DE CONJONCTURE

Trois notes de synthèse et un point de conjoncture pour suivre la situation et les perspectives à moyen terme de l'économie française. Le supplément "Séries longues" donne des tableaux et des graphiques sur 25 ans.

Abonnement 1 an (3 notes + 1 point + 1 supplément Séries longues)
France : 210 FF - Europe : 263 FF - Reste du monde : 309 FF
32,01 euros 40,09 euros 47,11 euros

CONJONCTURE IN FRANCE

Deux fois par an une synthèse de la conjoncture économique de la France rédigée en anglais.

Abonnement 1 an (2 numéros)
France : 50 FF - Europe : 63 FF - Reste du monde : 75 FF
7,62 euros 9,60 euros 11,43 euros

NOTE DE CONJONCTURE INTERNATIONALE
DIRECTION DE LA PRÉVISION

Deux fois par an, un panorama de la conjoncture mondiale dressé par la Direction de la Prévision. En supplément, deux points de conjoncture internationale.

Abonnement 1 an (2 notes + 2 points)
France : 155 FF - Europe : 194 FF - Reste du monde : 245 FF
23,63 euros 29,58 euros 37,35 euros

BULL [redacted] D'ABONN [redacted]

À RETOURNER À : INSEE - CNGP BP 2718 - 80027 AMIENS Cedex 01
Tél. : 03 22 92 73 22 - Fax : 03 22 97 92 95

Veuillez noter mon abonnement aux publications suivantes : ....

....

Nom ou raison sociale : ....

Activité : .... Tél : .... Fax : ....

Adresse : ....

....

Je règle un montant de ....FF * (total des abonnements) par : ☐ **chèque** (à l'ordre de l'Insee).

☐ **Carte bancaire.** ☐ Visa ☐ Mastercard ☐ Eurocard ***(seules cartes acceptées)***

Carte N° |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| Expire au : |_|_| |_|_|

Date : **Signature obligatoire :**

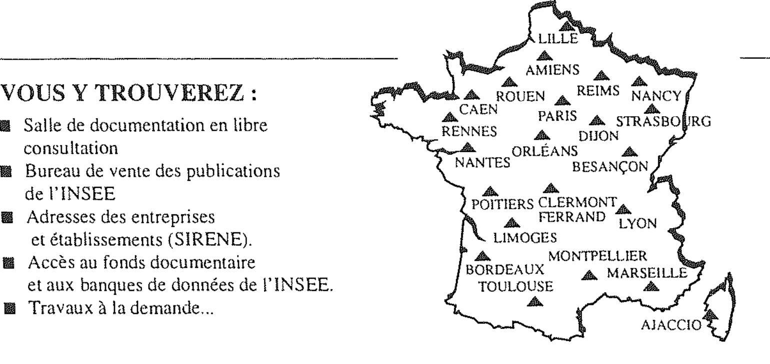
*pour l'Europe libellé en FF ou en euros.

N.B. : Toute commande par fax devra être obligatoirement confirmée par courrier.

L'INSEE DANS VOTRE RÉGION

VOUS Y TROUVEREZ :

- Salle de documentation en libre consultation
- Bureau de vente des publications de l'INSEE
- Adresses des entreprises et établissements (SIRENE).
- Accès au fonds documentaire et aux banques de données de l'INSEE.
- Travaux à la demande...



LE SERVICE INSEE 24H/2

08 36 68 07 60 (2,23 F/mn)

- indices
- informations
- adresses

et sur Minitel

36.15 INSEE (1,01F/mn)
36.17 INSEE *les informations directement chez vous par télécopie* (5,57F/mn)

ALSACE
Cité administrative GAUJOT
14, rue du Maréchal Juin,
67084 STRASBOURG CEDEX
Tél. : 03 88 52 40 40

AQUITAINE
33, rue de Saget,
33076 BORDEAUX CEDEX
Tél. : 05 57 95 04 00

AUVERGNE
3, place Charles de Gaulle, BP 120,
63403 CHAMALIERES CEDEX
Tél. : 04 73 31 82 00

BOURGOGNE
2, rue Hoche, BP 1509,
21035 DIJON CEDEX
Tél. : 03 80 40 67 48

BRETAGNE
"Le Colbert",
36, place du Colombier,
35082 RENNES CEDEX
Tél. : 02 99 29 33 33

CENTRE
43, avenue de Paris, BP 6719,
45067 ORLÉANS CEDEX 2
Tél. : 02 38 69 53 35

CHAMPAGNE-ARDENNE
10, rue Edouard Mignot,
51079 REIMS CEDEX
Tél. : 03 26 48 61 00

CORSE
Résidence Cardo,
rue des Magnolias,
BP 907,
20700 AJACCIO CEDEX 9
Tél. : 04 95 23 54 50

FRANCHE-COMTÉ
Immeuble "Le Major",
83, rue de Dôle,
BP 1997,
25020 BESANCON CEDEX
Tél. : 03 81 41 61 66

ILE-DE-FRANCE
INSEE Info Service,
accueil, librairie, consultation,
travaux sur mesure et sur rendez-vous
Tour "Gamma A",
195, rue de Bercy,
75582 PARIS CEDEX 12
Tél. : 01 41 17 66 11

Direction Régionale
7, rue Stephenson,
Montigny-le-Bretonneux
78188 ST-QUENTIN-EN-YVELINES CEDEX
Tél. : 01 30 96 90 99

LANGUEDOC-ROUSSILLON
274, allée Henri II de Montmorency,
"Le Polygone",
34064 MONTPELLIER CEDEX 2
Tél. : 04 67 15 71 11

LIMOUSIN
50, avenue Garibaldi,
87031 LIMOGES CEDEX
Tél. : 05 55 45 20 07

LORRAINE
15, rue du Général Hulot, BP 3846,
54029 NANCY CEDEX
Tél. : 03 83 91 85 85

MIDI-PYRÉNÉES
36, rue des 36 ponts,
31054 TOULOUSE CEDEX
Tél. : 05 61 36 61 13

NORD - PAS-DE-CALAIS
130, avenue du Président J.-F. Kennedy,
BP 769, 59034 LILLE CEDEX
Tél. : 03 20 62 86 33

BASSE-NORMANDIE
93-95, rue de Geôle,
14052 CAEN CEDEX
Tél. : 02 31 15 11 11

HAUTE-NORMANDIE
8, quai de la Bourse,
76037 ROUEN CEDEX
Tél. : 02 35 52 49 94

PAYS DE LA LOIRE
105, rue des Français Libres, BP 67401,
44274 NANTES CEDEX 02
Tél. : 02 40 41 79 80

PICARDIE
1, rue Vincent Auriol,
80040 AMIENS CEDEX 1
Tél. : 03 22 91 39 39

POITOU-CHARENTES
5, rue Sainte Catherine, BP 557
86020 POITIERS CEDEX
Tél. : 05 49 30 01 01

PROVENCE - ALPES - CÔTE D'AZUR
17, rue Menpenti,
13387 MARSEILLE CEDEX 10
Tél. : 04 91 17 59 50

RHÔNE-ALPES
165, rue Garibaldi, BP 3196,
69401 LYON CEDEX 03,
(Cité administrative de la Part-Dieu)
Tél. : 04 78 63 22 02

EN OUTRE - MER :

ANTILLES-GUYANE
Direction Inter-Régionale
41, rue Bébian
BP 300
97158 POINTE-A-PITRE CEDEX
Tél. : 0 590 21 47 07

GUADELOUPE
Service Régional
Rue Paul Lacavé, BP 96,
97102 BASSE-TERRE
Tél. : 0 590 99 36 36

GUYANE
Service Régional
Avenue Pasteur, BP 6017,
97306 CAYENNE CEDEX
Tél. : 0 594 31 61 00

MARTINIQUE
Service Régional, Centre Delgrès
Boulevard de la Pointe des Sables
Les Hauts de Dillon, BP 641
97262 FORT DE FRANCE CEDEX
Tél. : 0 596 60 73 60

RÉUNION
Direction Régionale,
15, rue de l'Ecole, BP 13,
97408 ST DENIS MESSAG CEDEX 9
Tél. : 0 262 48 89 21

INSEE - DIRECTION GÉNÉRALE
Unité Communication Externe
Timbre H501 - 18, bd Adolphe-Pinard
75675 Paris Cedex 14 - FRANCE

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Tél. renseignements : 01 41 17 66 11
Tél. administration : 01 41 17 50 50

INSEE MÉTHODES
N° 84 - 85 - 86

ACTES DES JOURNÉES DE MÉTHODOLOGIE STATISTIQUE

17 et 18 mars 1998



Ce volume rassemble les communications des VI-èmes « Journées de méthodologie statistique » qui se sont tenues à Paris les 17 et 18 mars 1998.

Les sujets abordés sont le panel européen « de ménages », la collecte et les enquêteurs, le logiciel de calcul de précision POULPE, les indices et les expériences de formation active aux enquêtes.

Ces journées sont une occasion d'échange de savoirs et d'expériences entre statisticiens de l'Insee ou du système statistique public. Elles sont aussi ouvertes aux universitaires et experts français et étrangers travaillant dans le domaine statistique. Ainsi a-t-on pu apprécier des interventions sur l'harmonisation des méthodes au niveau européen (Eurostat), le prochain recensement aux États-Unis (Bureau of the Census) et sur l'imputation des non-réponses (Statistique Canada).



ISSN 1142 - 3080
ISBN 2-11-066998 5
IMETO84
Mars 1999 - Prix : 228 F 34,76 €